Journal of Obsessive-Compulsive and Related Disorders 20 (2019) 39–49

Contents lists available at ScienceDirect

Journal of Obsessive-Compulsive and Related Disorders

journal homepage: www.elsevier.com/locate/jocrd

OCD-like checking in the lab: A meta-analysis and improvement of an experimental paradigm

Marcel A. van den Hout*, Eva A.M. van Dis, Clair van Woudenberg, Ilse H. van de Groep

Department of Clinical Psychology, Utrecht University, The Netherlands

ARTICLE INFO

Keywords:
OCD
Repeated checking
Memory
Meta-analysis
Experiment
Paradigm

ABSTRACT

Van den Hout and Kindt (2003a) developed a *Virtual Gas Stove Checking* paradigm. They demonstrated that repeated checking resulted in lower confidence and reduced the vividness and detail of recollections. Over the past decades, many experiments have used (an adaptation of) this experimental paradigm to study phenomena related to obsessive compulsive disorders (OCD). The first aim of the present study was to conduct a meta-analysis of experiments ($k = 28$; $N = 1662$) on the repeated checking paradigm. Repeated checking was found to have large effects on decreases in memory confidence, vividness and detail. Unexpectedly, repeated checking also produced small reductions in memory accuracy. The second aim of the present study was to develop an improved version of the checking paradigm in which 1) stimuli presentations were fully balanced; and 2) the checking latency was comparable across stimuli in order to 3) assess actual checking behavior. The improved version (Virtual checking task 2.0) replicated earlier findings on meta-memory.

1. General introduction

Although obsessive compulsive disorder (OCD) is a heterogeneous disorder, there are two prominent features: 1) patients tend to be uncertain about cognitive functions like memory ("did I really shut the door?"; Hermans, Martens, De Cort, Pieters, & Eelen, 2003; MacDonald, Antony, McLeod & Richter, 1997; Dar, Rish, Hermesh, Taub, & Fux, 2000; Hermans et al., 2008), and 2) around 80% patients engage in repetitive checking (Ruscio, Stein, Chiu, & Kessler, 2010). Patients typically maintain that checking serves to reduce or prevent cognitive uncertainty (Tallis, 1995). Around the turn of the century, several authors suggested that although repetitive checking may indeed be motivated by the wish to reduce uncertainty, it paradoxically increases rather than reduces uncertainty about checked issues (Rachman, 2002; Salkovskis & Forrester, 2002; Tolin et al., 2001; van den Hout & Kindt, 2003a). In their 2003a paper, van den Hout and Kindt reported three experiments that critically tested this hypothesis. Healthy volunteers engaged in a *virtual gas stove* checking task and were asked to turn on, turn off and check 3 out of 6 gas rings by turning corresponding knobs with a computer mouse. After a first checking trial (the pre-test), participants were tested for memory accuracy (which rings did you have to check?) and rated their confidence in the memory accuracy, as well as the vividness and detail of their memory. Ratings were scored on visual analogue scales (VASes). Subsequently, half of the participants (i.e., relevant checking group) this process with various configurations of the gas rings. After the twentieth trial, they completed a post-test, which was identical to the pre-test. The other half of the group (i.e., irrelevant checking group) performed the same pre-test and post-test with the gas rings, however irrelevant stimuli (i.e., virtual light bulbs) were displayed instead of gas rings. Accuracy was good at pre-test and remained so at post-test in both conditions. Moreover, in line with the hypothesis, scores on memory confidence, vividness and detail dropped substantially in the relevant checking condition, whereas no such effect occurred in the irrelevant checking condition. This points to a psychological cascade that helps to understand the maintenance of compulsive checking. The authors argued that the repetition of the gas ring checking increased the familiarity of the checked stimuli, rendering the checking an automatic routine. The latter implies a reduction in perceptual processing, culminating in reduced ratings for confidence, vividness, and detail at post-test. In the irrelevant checking group, the post-test stimuli were relatively new, explaining why no such effects were observed here. Therefore, uncertainty may promote checking, yet checking backfires and may ironically serve to enhance uncertainty.

The virtual checking task provides an experimental model of the effects of perseverative, OCD-like checking. The original publication (van den Hout & Kindt, 2003a) was relatively well-cited and the task was used in several laboratories. The first aim of the present paper was to evaluate the robustness of the task effects by undertaking a brief meta-analysis. There is another aim. The checking paradigm as described above has at least three shortcomings that call for an improved

* Correspondence to: Heidelberglaan 1, 3584 CS Utrecht, The Netherlands.
E-mail address: m.vandenhout@uu.nl (M.A. van den Hout).

https://doi.org/10.1016/j.jocrd.2017.11.006
Received 31 July 2017; Received in revised form 8 November 2017; Accepted 28 November 2017
Available online 12 December 2017

version. First, the crucial dependent variables (memory confidence for the last check and the vividness and detail of the recollection) are *self-reports*, rated on VASes. While, arguably, self-reports are appropriate to study experiential phenomena of the present type (van den Hout, Engelhard, & McNally, 2017), these are silent about actual behavior and processes that cannot be assessed via self-report. For instance, it has been suggested (van den Hout & Kindt, 2003a) that, as checking continues, it becomes automated. If so, with the repetition it should gradually take less effort, which should surface in faster checking. Ideally then, the paradigm should monitor participants' actual checking behavior, to enable testing for whether checking automatization really occurs. In addition, researchers could study other behavioral parameters, such as the numbers and types of errors occurring or spontaneous re-checking. The original paradigm does not allow for this and a new, upgraded version should. Second, *condition* and *stimulus-materials* were confounded and unbalanced in the original paradigm. The experiment has three phases: pre-test→ repeated checking→ post-test. In the relevant checking condition, the stimuli were Gas rings→ Gas rings→ Gas rings. In the irrelevant checking condition, the stimuli were Gas rings→ Light bulbs→ Gas rings. To have a properly balanced

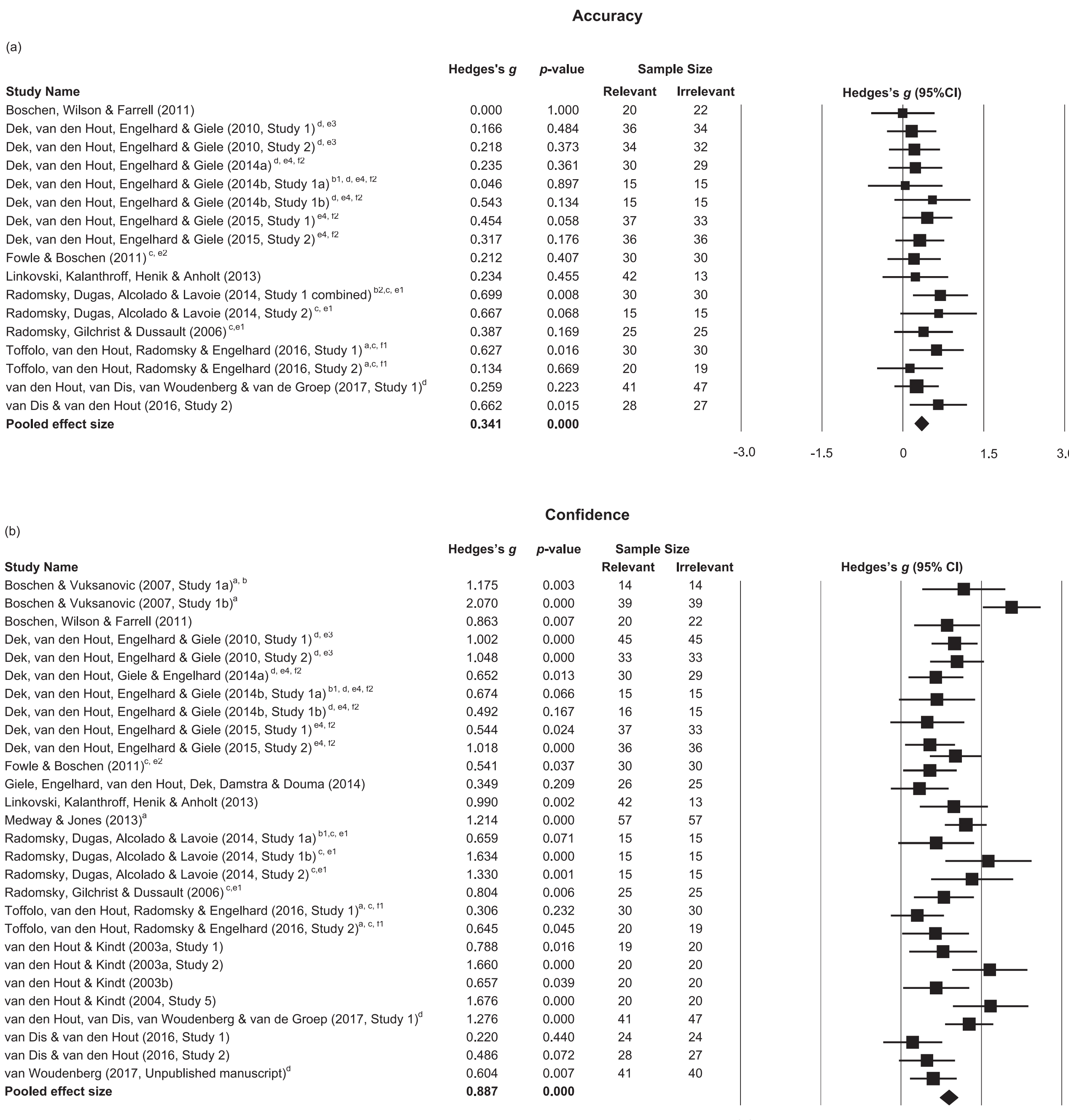
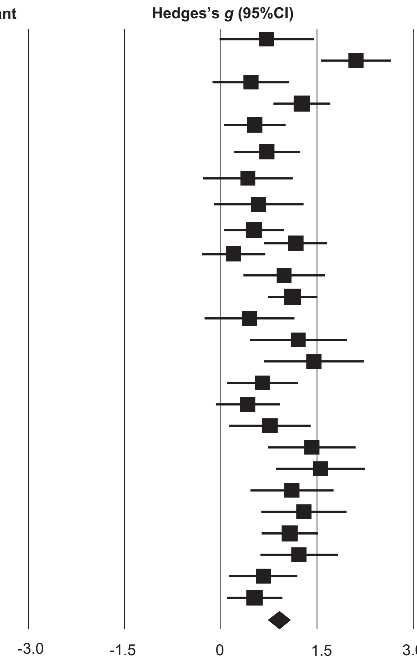


Fig. 1. Forest plot of the effect sizes (Hedges's *g*) for all studies reporting a Time (pre- vs. post-test) x Revelance (relevant vs. irrelevant checking) interaction on the outcome variables (a) accuracy, (b) confidence, (c) vividness and (d) detail, using a random effects model. CI = confidence interval. Subscripts denote how the reported studies differ from the original design (cf. van den Hout & Kindt, 2003a); a = within subjects design instead of a mixed design; b1 = OCD patients instead of healthy subjects, b2 combination of OCD and healthy subjects; c = non-virtual checking instead of virtual checking; d = fully balanced design instead of a partially balanced design; the subscript e denotes the use of different stimuli that gas rings vs. light bulbs, with e1 = stove vs. sink, e2 = colored plates vs. colored bowls, e3 = large green circles vs. small grey circles, e4 = large circles with star vs. small grey circles; the subscript f indicates the use of a different number of trials than 20, with f1 = 14 trials and f2 = 15 trials (Boschen et al., 2011; Giele et al., 2014; Linkovski et al., 2013; Medway and Jones, 2013; Radomsky et al., 2006; Radomsky et al., 2014; Toffolo et al., 2016; van den Hout and Kindt, 2003b; van den, 2016).

Vividness

(c)

Study Name	Hedges's *g*	*p*-value	Sample Size Relevant	Sample Size Irrelevant
Boschen & Vuksanovic (2007, Study 1a)[a, b]	0.719	0.058	14	14
Boschen & Vuksanovic (2007, Study 1b)[a]	2.112	0.000	39	39
Boschen, Wilson & Farrell (2011)	0.473	0.124	20	22
Dek, van den Hout, Engelhard & Giele (2010, Study 1)[d, e3]	1.266	0.000	45	45
Dek, van den Hout, Engelhard & Giele (2010, Study 2)[d, e3]	0.531	0.032	33	33
Dek, van den Hout, Giele & Engelhard (2014a)[d, e4, f2]	0.723	0.006	30	29
Dek, van den Hout, Engelhard & Giele (2014b, Study 1a)[b1, d, e4, f2]	0.427	0.235	15	15
Dek, van den Hout, Engelhard & Giele (2014b, Study 1b)[d, e4, f2]	0.595	0.097	16	15
Dek, van den Hout, Engelhard & Giele (2015, Study 1)[e4, f2]	0.519	0.031	37	33
Dek, van den Hout, Engelhard & Giele (2015, Study 2)[e4, f2]	1.170	0.000	36	36
Fowle & Boschen (2011)[c, e2]	0.201	0.432	30	30
Linkovski, Kalanthroff, Henik & Anholt (2013)	0.991	0.002	42	13
Medway & Jones (2013)[a]	1.121	0.000	57	57
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 1a)[b1,c, e1]	0.452	0.210	15	15
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 1b)[c, e1]	1.211	0.002	15	15
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 2)[c, e1]	1.456	0.000	15	15
Radomsky, Gilchrist & Dussault (2006)[c, e1]	0.651	0.023	25	25
Toffolo, van den Hout, Radomsky & Engelhard (2016, Study 1)[a, c, f1]	0.424	0.100	30	30
Toffolo, van den Hout, Radomsky & Engelhard (2016, Study 2)[a, c, f1]	0.770	0.018	20	19
van den Hout & Kindt (2003a, Study 1)	1.421	0.000	19	20
van den Hout & Kindt (2003a, Study 2)	1.556	0.000	20	20
van den Hout & Kindt (2003b)	1.113	0.001	20	20
van den Hout & Kindt (2004, Study 5)	1.299	0.000	20	20
van den Hout, van Dis, van Woudenberg & van de Groep (2017, Study 1)[d]	1.078	0.000	41	47
van Dis & van den Hout (2016, Study 1)	1.223	0.000	24	24
van Dis & van den Hout (2016, Study 2)	0.667	0.015	28	27
van Woudenberg (2017, Unpublished manuscript)[d]	0.531	0.018	41	40
Pooled effect size	**0.902**	**0.000**		



Detail

(d)

Study name	Hedges's *g*	*p*-value	Sample size Relevant	Sample size Irrelevant
Boschen & Vuksanovic (2007, Study 1a)[a, b]	0.735	0.053	14	14
Boschen & Vuksanovic (2007, Study 1b)[a]	2.098	0.000	39	39
Boschen, Wilson & Farrell (2011)	0.558	0.071	20	22
Dek, van den Hout, Engelhard & Giele (2010, Study 1)[d, e3]	1.352	0.000	45	45
Dek, van den Hout, Engelhard & Giele (2010, Study 2)[d, e3]	0.622	0.013	33	33
Dek, van den Hout, Giele & Engelhard (2014a)[d, e4, f2]	0.822	0.002	30	29
Dek, van den Hout, Engelhard & Giele (2014b, Study 1a)[b1, d, e4, f2]	0.517	0.152	15	15
Dek, van den Hout, Engelhard & Giele (2014b, Study 1b)[d, e4, f2]	0.574	0.108	16	15
Dek, van den Hout, Engelhard & Giele (2015, Study 1)[e4, f2]	0.493	0.040	37	33
Dek, van den Hout, Engelhard & Giele (2015, Study 2)[e4, f2]	1.115	0.000	36	36
Fowle & Boschen (2011)[c, e2]	0.363	0.158	30	30
Linkovski, Kalanthroff, Henik & Anholt (2013)	0.789	0.014	42	13
Medway & Jones (2013)[a]	1.149	0.000	57	57
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 1a)[b1,c, e1]	0.938	0.012	15	15
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 1b)[c, e1]	0.973	0.010	15	15
Radomsky, Dugas, Alcolado & Lavoie (2014, Study 2)[c, e1]	1.638	0.000	15	15
Radomsky, Gilchrist & Dussault (2006)[c,e1]	0.691	0.016	25	25
Toffolo, van den Hout, Radomsky & Engelhard (2016, Study 1)[a, c, f1]	0.083	0.744	30	30
Toffolo, van den Hout, Radomsky & Engelhard (2016, Study 2)[a, c, f1]	1.014	0.002	20	19
van den Hout & Kindt (2003a, Study 1)	1.078	0.001	19	20
van den Hout & Kindt (2003a, Study 2)	1.046	0.002	20	20
van den Hout & Kindt (2003b)	0.720	0.025	20	20
van den Hout & Kindt (2004, Study 5)	1.330	0.000	20	20
van den Hout, van Dis, van Woudenberg & van de Groep (2017, Study 1)[d]	0.755	0.001	41	47
van Dis & van den Hout (2016, Study 1)	1.253	0.000	24	24
van Dis & van den Hout (2016, Study 2)	0.642	0.019	28	27
van Woudenberg (2017, Unpublished manuscript)[d]	0.553	0.014	41	40
Pooled effect size	**0.872**	**0.000**		

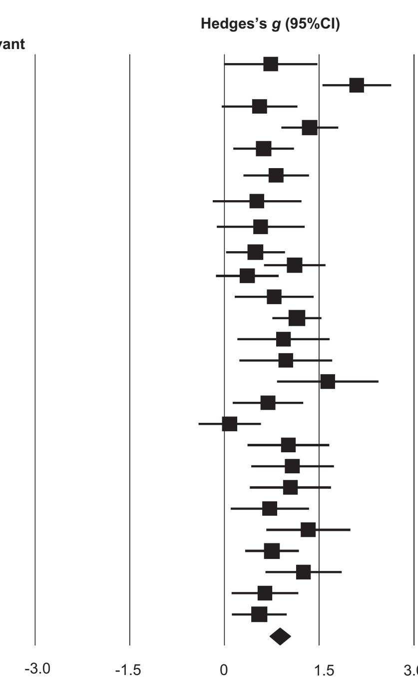


Fig. 1. (*continued*)

paradigm, there should be another relevant checking and another irrelevant checking condition. The relevant checking being: Light bulbs→ Light bulbs→ Light bulbs and the other irrelevant checking condition being: Light bulbs→ Gas rings→ Light bulbs. It may be noted that in two previous experiments, the gas rings and light bulbs were replaced by abstract objects (circles with different sizes and colors), and the design was fully balanced/not confounded. These experiments yielded the same meta-memory effects: drops in confidence, vividness, and detail in the relevant checking conditions, but not in the irrelevant conditions (Dek, van den Hout, Giele & Engelhard, 2010). A new program for the virtual checking experiment should by default use the fully balanced design. As a third shortcoming, it was assumed that both tasks (handling gas rings vs. light bulbs) needed comparable effort to carry out and that any minor differences would not affect the meta-memory discussed above: confidence, vividness, and detail. This should be evident from a) comparable response latencies developing in and across the task trials, and b) identical patterns of meta-memory data for the two tasks. However, the paradigm did not allow for verification of this assumption. Given that the phenomenon at hand is believed to result from automatization, it is particularly important (1) that the tasks are comparable in the amount of effort it takes to complete them and (2) to show the trial-by-trial development of the hypothetical automatization for both tasks. In short, then, the first aim of the present paper was to provide a meta-analytic overview of the published checking-paradigm studies. The second aim was to develop an adapted version of the paradigm to overcome the aforementioned problems: (1) it should assess actual checking behavior, (2) stimuli should be fully balanced, and (3) the experimental tasks should be comparable in terms of effort, measured in response latency over the various trials, while any minor differences would be immaterial to the cognitive measures.

2. Meta-analysis

2.1. Literature search

Relevant studies (published until June 9, 2017) were identified by systematically searching several electronical databases (i.e., PsychInfo, Pubmed, Embase and OpenGrey). Our search strategies included a combination of the terms memory, repeat*, persever* and check* (see Appendix A for exact search strategies). The electronic database search was supplemented by examining all papers that cited the original van den Hout & Kindt (2003a) study. We knew of one unpublished study from our laboratory and added this study to the meta-analysis as well.

2.1.1. Inclusion criteria

Studies were selected if they included: (1) (an adapted version of) the repeated checking task first described by van den Hout & Kindt (2003a); (2) (at least) one data point before and after the intervention (i.e., pre-test and post-test); (3) a direct comparison between repeated relevant checking and irrelevant checking. We also included studies with minor adaptations to the original design, such as (1) within-subject design instead of a mixed design, (2) inclusion of (both healthy controls and) OCD patients, (3) non-virtual checking instead of virtual checking, (4) a fully balanced design instead of a partially balanced design, (5) different stimuli and (6) different number of trials. For a more detailed description and overview of the adaptations for each included study, see Fig. 1.

2.1.2. Study selection

The search strategy, combined with the additional citation search, resulted in 364 unique articles which were independently screened by title and abstract by IG and ED using Covidence systematic review software (available at www.covidence.org). Disagreements ($k = 1$) were resolved through discussion. Of these articles, 44 full-texts were further assessed for eligibility. If the reported data were insufficient for effect size calculation, authors were contacted by IG and requested to provide the missing information ($k = 5$, response rate = 80%). In total, 19 articles were included, reporting 28 studies. All these studies were conducted on different samples of subjects, so the assumption of independence (Lipsey & Wilson, 2001) has presumably not been violated in our analysis.

2.1.3. Outcomes

The primary outcomes were the accuracy, confidence, vividness, and detail of ratings related to the participant's recollection of the last checking operation. We only considered interaction effects between Condition (Relevant vs. Irrelevant) × Time (Pre- vs. Post-test) for these outcome measures (corresponding means, standard deviations, p-values, F-tests, t-tests, or estimates of effect sizes). Studies that only reported accuracy in counts (i.e., number of mistakes made at test) rather than the Condition × Time interaction were excluded from analyses on accuracy. Reliable estimates of pre-post correlations are crucial (Cuijpers, Weitz, Cristea, & Twisk, 2016), hence we estimated these for each outcome measure based on data of eleven studies: accuracy: $r = .089$, $p = .054$; confidence: $r = .205$, $p < .001$; vividness: $r = .280$, $p < .001$; detail: $r = .382$, $p < .001$ (van den Hout & Kindt, 2004 [Study 5]; Boschen & Vuksanovic, 2007; Dek, van den Hout, Giele & Engelhard, 2010 [Study 1 and 2]; Dek, van den Hout, Giele & Engelhard, 2014a, 2014b [Study 1 and 2]; Dek, van den Hout, Engelhard, & Giele, Cath, 2015; van Dis & van den Hout, 2016 [Study 1 and 2], and the data reported below).

2.1.4. Data analysis

All analyses were conducted using the Comprehensive Meta-analysis software (Version 3.3.070), using a random effects model. Hedges' g was used for effect size estimation, given its ability to control for variations in sample sizes between studies (Borenstein, Hedges, Higgins, & Rothstein, 2009). We used the I^2 statistic as indicator of heterogeneity, which displays the proportion of the observed variance reflecting variance in true effect sizes rather than sampling error (Borenstein et al., 2009). I^2 ranges from 0% to 100%, where low values mean that most of the dispersion of effects would disappear if the sampling error could be removed (and vice versa, high values mean that most of the observed dispersion would remain; Borenstein, Higgins, Hedges, & Rothstein, 2017). Besides the I^2 statistic, we also calculated prediction intervals to estimate the absolute range of effects across populations for each outcome measure (see Borenstein et al., 2017). Risk of publication bias was tested using Egger's test (one-tailed; Egger, Smith, Schneider, & Minder, 1997). In addition, we assessed risk of publication bias through trim and fill technique (Duval & Tweedie, 2000), which is a funnel-plot-based technique of testing and adjusting for publication bias.

2.2. Results

2.2.1. Memory accuracy, confidence, vividness, and detail

On memory accuracy, 17 studies ($N = 1112$; 63 OCD patients) reported an interaction effect between Condition (relevant vs. irrelevant checking) and Time (pre- vs. posttest). As can be seen in Fig. 1a, these studies demonstrated that repeated checking has a small effect on memory accuracy, $g = .341$, 95% CI [.212, .469]. For memory confidence, we pooled effects of 28 studies ($N = 1622$; 93 OCD patients) demonstrating that repeated checking has a large effect on memory confidence, $g = .887$, 95% CI [.720, 1.054] (Fig. 1b). A similar, large effect of repeated checking was observed on memory vividness (27 studies, $N = 1568$; 93 OCD patients), $g = .902$, 95% CI [.733, 1.071] (see Fig. 1c), and on detail (27 studies, $N = 1568$; 93 OCD patients), $g = .872$, 95% CI [.709, 1.034] (see Fig. 1d).

2.2.2. Sensitivity analyses

We conducted a sensitivity analysis for each outcome, to assess the robustness of our results. For each outcome, we repeatedly calculated the pooled effect size while leaving out one study, to see whether one study strongly influenced the pooled effect. These analyses showed that the effect size did not fundamentally change for any of the outcome measures, proving the robustness of our results. When leaving out the most extreme study, the effect size of Confidence (.887) either changed to .836 or .912; Vividness (.902) to .848 or .932; Detail (.872) to .817 or .905; and Accuracy (.341) to .317 or .358.

2.2.3. Publication bias

Egger's test of the intercept suggests there is no publication bias for accuracy, $t(15) < 1$, $p = .405$, confidence, $t(26) < 1$, $p = .247$, vividness, $t(25) < 1$, $p = .266$, and detail, $t(25) < 1$, $p = .264$. Duval and Tweedie's trim and fill procedure revealed that no studies had to be imputed for any of the outcome variables. Hence, both tests indicate there is a low risk of publication bias for all outcome measures. Fig. C1 shows the funnel plots with standard error by Hedges' g for all outcome measures (see Appendix C).

2.2.4. Test for heterogeneity

Our results indicated that heterogeneity for memory accuracy was probably small, I^2 = .00. Estimates of absolute variation indicate that the estimated effect of accuracy (Hedges' *g*) may vary across populations from .202 to .480 (i.e., prediction interval). For memory confidence, vividness, and detail heterogeneity may be moderate, with I^2 = 58.29, I^2 = 57.89, and I^2 = 54.72 respectively. The prediction intervals range from .171 to 1.603 for confidence; from .187 to 1.617 for vividness and .205–1.539 for detail. As we did not pre-specify any subgroup differences in advance, we decided not to further investigate explanations for heterogeneity (see Higgins & Green, 2011), but added an explorative analysis of heterogeneity in different subgroups in Appendix C.

2.3. Discussion meta-analysis

The meta-analysis demonstrated the meta-memory findings are robust. Relative to irrelevant checking, relevant checking induced large and reliable drops in memory confidence, vividness, and detail. Moreover, the meta-analysis revealed a small decline in accuracy in the relevant checking condition in comparison to the irrelevant checking condition. Interestingly, such an effect was not detected in most of the individual studies we reviewed. In the general discussion, we will elaborate on this discrepancy and provide a critical discussion of the main findings.

3. Virtual checking task 2.0

3.1. Introduction

The aim was to improve the original task by allowing for assessment, trial by trial, of the nature and timing of the behavioral responses. In both the original and the present version of the paradigm, there were two stimulus conditions: virtual gas rings and virtual light bulbs. We wanted the response latencies for both stimuli, and their development over time, to be comparable. We attempted to reach this equivalence by small variations in the display of the rings/bulbs and in the way participants could interact with the rings/lights. Pilot experiments proved latency-equivalence for the two tasks to be a sensitive issue: minor variations made latencies for one of the two tasks conditions slower or faster than the other. However, in line with the results from the meta-analysis given above, the paradigm proved robust: despite between-task differences in latencies, decreases in vividness, detail and confidence occurred in the relevant checking condition, relative to the irrelevant checking condition. Based on these pilots, we settled for the paradigm that is described below.

3.2. Method

3.2.1. Participants

The study sample included 88 students who were recruited at Utrecht University from June to December 2016 (*M* age = 22, *SD* = 2.7, 63 females). Participants were remunerated by course credits or small financial reward and all provided written informed consent prior to participation.

3.2.2. Task

Participants performed an adapted version of the virtual checking task (van den Hout & Kindt, 2003a) programmed in Matlab, R2015b, which has been made available online (see Supplementary Material). The computer task included a virtual six-burner stove and a set of six light bulbs that participants could turn on and off using the computer mouse (see Fig. 2). Each trial started with a 4 s (s) presentation of a schematic diagram that presented three random screen positions of gas rings or light bulbs that needed to be turned on. After turning on the gas rings or light bulbs, participants were asked to turn off these stimuli and finally to check whether they were turned off correctly. The task took about 15 min to complete.

3.2.3. Procedure

Participants practised with one gas ring and one light bulb trial, and then continued with the first trial (either a gas ring or light bulb trial). After the first trial, participants filled out the questions about memory accuracy, confidence, vividness and detail of that first checking trial (i.e., pre-test). Next, participants performed 20 checking trials that were either all similar to the first trial (i.e., relevant checking group) or different from the first trial (i.e., irrelevant checking group). After the 20 checking trials, participants performed the post-test, using the same stimulus type they encountered earlier in the pre-test, and subsequently rated memory accuracy, confidence, vividness and detail of that final checking bout. Importantly, the stimulus types were fully balanced across conditions. Therefore, participants in the relevant checking condition were either (1) checking gas rings during all phases (GGG) or (2) checking light bulbs during all phases (LLL). In contrast, participants in the irrelevant checking condition were either (1) checking the gas rings in the pre-and post-test, and the light bulbs during the checking trials (GLG) or (2) checking the light bulbs during the pre-and post-test, and the gas rings during the checking trials (LGL).

3.2.4. Assessments

3.2.4.1. Accuracy. During the pre- and post-test, participants were presented with a schematic depiction of either six gas rings or light bulbs and asked to indicate which gas rings or light bulbs they had been instructed to turn on. In previous studies authors had assessed accuracy by comparing the items to be checked with the participants' recollection (i.e., "subjective accuracy"). Any mismatch between the instruction (e.g., "Check gas rings 2, 4 and 6″) and self-reported checks (e.g., a participant indicated having checked gas rings 2, 4 and 5) was considered as inaccurate subjective accuracy (dichotomous variable). The present paradigm also allowed for assessing whether the items included in the instruction were actually the ones that were checked (i.e., "objective accuracy"). Any mismatch between the instruction (e.g., "Check gas rings 2, 4 and 6″) and actual checks (e.g., a participant checked gas rings 2, 4 and 5) was considered as inaccurate objective accuracy (dichotomous variable).

3.2.4.2. Memory confidence, vividness and detail. Participants were asked to indicate on a VAS (ranging from 0 = *absolutely not confident* to 100 = *absolutely confident*) how confident they were that they answered the accuracy question correctly. Moreover, they were asked to rate the vividness and detail of their recollection of the last checking trial they performed on two VASes, ranging from 0 = *not vivid* to 100 = *extremely vivid*, and 0 = *not detailed* to 100 = *extremely detailed*, respectively.

3.3. Results

3.3.1. Memory effects: accuracy, confidence, vividness and detail

In terms of *accuracy*, Table 1 demonstrates that participants were relatively accurate in remembering which items they were supposed to check, with subjective accuracy scores of around 85%. In contrast, objective accuracy was lower, with accuracy scores ranging from 41% to 46%. Two separate 2 × 2 × 2 mixed analysis of variance (ANOVAs) with Time (pre-test vs. post-test) as the within group factor, and Relevance (relevant checking vs. irrelevant checking) and Stimulus (gas vs. lights) as between group factors, tested whether repeated checking affected self-reported and behavioral accuracy over time. Both ANOVAs showed no Time × Relevance × Stimulus interaction effects for subjective accuracy, $F(1, 84) < 1$, $p = .537$, $\eta_p^2 = .01$, and objective accuracy, $F(1, 84) = 1.09$, $p = .300$, $\eta_p^2 = .01$, nor for Time × Relevance, $F(1, 86) = 2.86$, $p = .094$, $\eta_p^2 = .03$ and $F(1, 86) = 1.81$, $p = .182$, $\eta_p^2 = .02$, respectively.

The VAS scores on memory confidence, detail and vividness were analyzed by three separate 2 × 2 × 2 mixed ANOVAs with Time (pre-test vs. post-test) as the within group factor, and Relevance (relevant

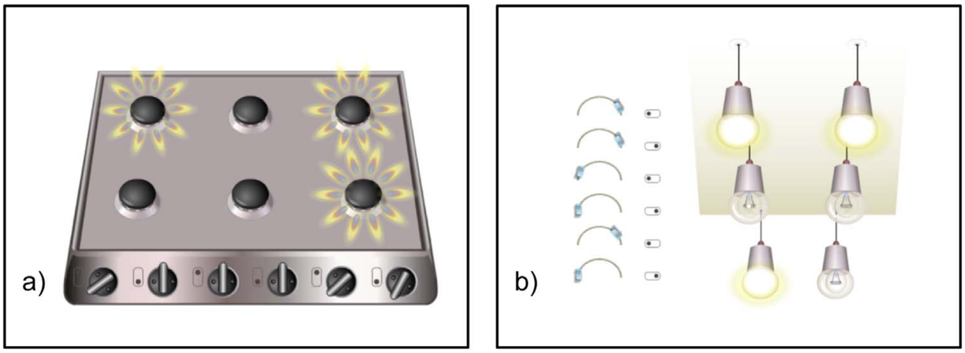


Fig. 2. Experimental stimuli of the adapted virtual checking task included a) gas rings and b) light bulbs.

Table 1
Percentage of participants that accurately reported or behaved conform instruction.

		Relevant	Irrelevant
Subjective	*Pre*	88%	85%
	Post	78%	94%
Objective	*Pre*	46%	43%
	Post	41%	45%

checking vs. irrelevant checking) and Stimulus (gas vs. lights) as between group factors. In terms of *confidence*, Fig. 3 (left panel) clearly shows that irrespective of stimulus type, confidence decreased only after relevant checking, and not after irrelevant checking. In line with this observation, the overall interaction between Time × Relevance × Stimulus was not statistically significant, $F(1, 84) < 1$, $p = .767$, $\eta_p^2 = .00$, yet crucially, we found a strong Time × Relevance interaction, $F(1, 86) = 33.50$, $p < .001$, $\eta_p^2 = .28$, reflecting a decrease in confidence in the relevant checking group, $t(40) = 7.35$, $p < .001$, while no difference over time was found in the irrelevant checking group, $t(46) < 1$, $p = .743$. The effects on *vividness* were completely consonant with the findings on confidence (see Fig. 3, central panel). Again, the Time × Relevance × Stimulus was not statistically significant, $F(1, 84) < 1$, $p = .501$, $\eta_p^2 = .00$, while we found again a strong Time × Relevance interaction, $F(1, 86) = 27.90$, $p < .001$, $\eta_p^2 = .25$. In line with our expectations, vividness ratings decreased over time after relevant checking, $t(40) = 6.45$, $p < .001$, but did not change over time after irrelevant checking, $t(46) < 1$, $p = .961$. In terms of *detail*, there were similar patterns (see Fig. 3, right panel). The Time × Relevance × Stimulus interaction was again not statistically significant, $F(1, 84) < 1$, $p = .833$, $\eta_p^2 = .00$, in contrast to the Time × Relevance interaction, $F(1, 86) = 15.96$, $p < .001$, $\eta_p^2 = .16$. The relevant checkers reported a decrease in memory detail over time, $t(40) = 3.91$, $p < .001$, while the detail ratings did not change over time for irrelevant checkers, $t(46) =$ 1.41, $p = .166$.

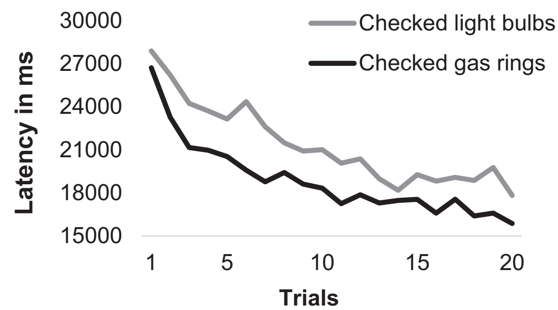


Fig. 4. Total duration (in ms) per checking trial (i.e., turning on, off and checking) for 20 checks.

3.3.2. Latency

The response latencies and their development over time are displayed in Fig. 4 for both the gas-checking and light bulb checking conditions. Response latencies were subjected to a 20 × 2 × 2 mixed ANOVA comparing Time (checking trials 1–20) as within group factor, while Relevance (relevant checking vs. irrelevant checking) and Stimulus (gas rings vs. light bulbs) served as between group factors. The sphericity assumption was violated for Time, hence we used Greenhouse-Geisser corrections. Fig. 4 shows considerable speeding up over trials, which was reflected in a significant main effect for Time $F(11.41, 958.29) = 37.62$, $p < .001$, $\eta_p^2 = .31$. Further trend analyses showed that the effect of Time followed a linear, $F(1, 84) = 231.81$, $p < .001$, $\eta_p^2 = .73$, and quadratic trend, $F(1, 84) = 85.38$, $p < .001$, $\eta_p^2 = .50$. In addition, Fig. 4 seems to suggest differences between stimuli: overall, responses to gas stimuli tended to be faster than to lights. Indeed, there was a significant main effect of Stimulus, $F(1, 84) = 5.61$, $p = .020$, η_p^2

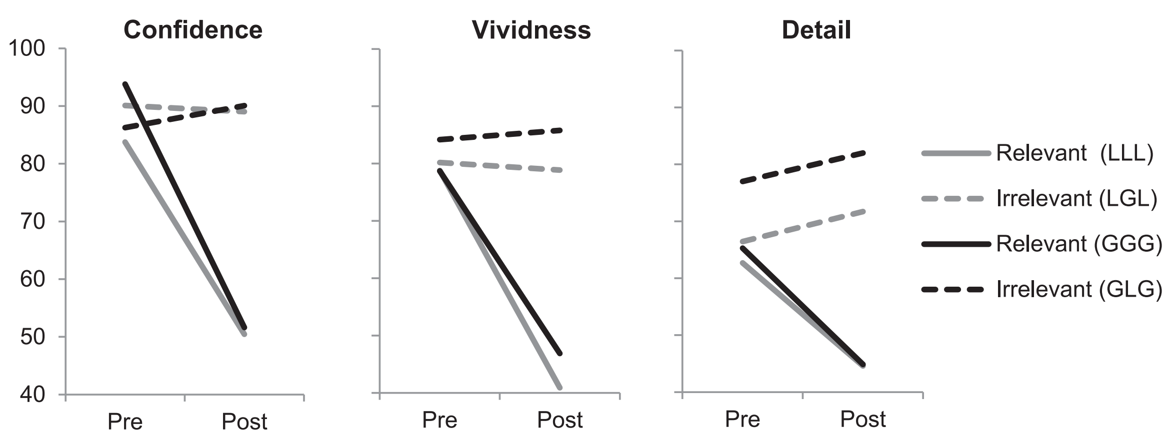


Fig. 3. Reported memory confidence, vividness and detail before and after repeated checking. LLL = light bulb trials during pre, checking, post; LGL = light bulb trials during pre and post, gas ring trials during checking; GGG = gas ring trials during pre, checking, post; LGL = gas ring trials during pre and post, light bulb trials during checking.

= .06. Furthermore, the development over time was comparable with the two stimuli: there was no interaction of Time × Stimulus, $F(11.41, 958.29) < 1$, $p = .409$, $\eta_p^2 = .01$, nor of Time × Stimulus × Condition, $F(19, 86) < 1$, $p = .759$, $\eta_p^2 = .01$.

3.4. Discussion virtual checking task 2.0

The current experiment used an updated version of the virtual checking task and found similar results. That is, compared to irrelevant checking, relevant checking induced decreases in memory confidence, vividness and detail. We found no effects of relevant checking on accuracy. Nevertheless, the study revealed unanticipated differences between "subjective accuracy" (i.e., match/mismatches between instructed responses and remembered responses) and 'objective accuracy" (i.e., match/mismatches between actual responses and remembered responses). Participants were considerably more accurate on subjective accuracy than on objective accuracy. Interestingly, we also found that checking latencies reduced over time (irrespective of condition), with a strong acceleration in the beginning, leveling off after about 10 trials.

Our results are in line with earlier findings obtained from an experimental design that was not fully balanced and, strictly speaking, contained, an experimental confound (see introduction). Likewise, earlier versions of the task did not match the conditions in terms of effort needed to complete the task, nor allowed for the assessment of actual behavior: speed of responding on the various trials and discrepancies between actual and remembered responses. In our updated checking task, the confound was removed and the tasks were highly comparable in terms of effort (for a critical and more detailed account, see general discussion).

Given the similarity between previous findings and the results of our present study, it is extremely unlikely that earlier findings were affected by the limited balancing/confound. Likewise, our replication of earlier findings makes it unlikely that these previous results were affected by between-condition differences in required effort. It should be noted that we were unable to develop tasks that were exactly similar in terms of required effort, as measured by response latencies. However, there was no indication that differences in experimental load affected the results.

4. General discussion

We aimed to evaluate the robustness of an experimental model of compulsive checking by conducting a meta-analysis of published studies. We also intended to improve the paradigm by (1) overcoming a potential confound in the earlier version, (2) adding behavioral assessments to complement the self-reports used earlier and (3) rendering the two paradigm tasks comparable in terms of response latencies.

4.1. Meta-analysis

The meta-analysis yielded clear results. Scores dropped reliably in the experimental condition on the variables of primary interest (confidence, vividness and detail) relative to the control condition of irrelevant checking, with effect sizes exceeding $g = .875$. The study designs were highly comparable with deviations in terms of characteristics of the participants, number of trials, nature of the stimuli being too infrequent to allow for statistical comparisons (however, see Appendix B for explorative subgroup comparisons).

For reasons that go beyond the scope of this paper, data collected on accuracy in the various papers (see also 3.3.1) are hard to handle statistically. Researchers have chosen different options, including just presenting the numbers (with no testing for statistical significance) or testing it statistically with a 2 × 2 mixed ANOVA. We pooled these interaction effects and unexpectedly found a small effect for repeated checking on memory accuracy, indicating that repeated checking slightly reduced memory accuracy. Note that almost all individual studies did not find an accuracy effect, illustrating the importance of conducting a meta-analysis, given that individual studies are often underpowered to detect small effects (Borenstein et al., 2009). Given that the pooled effect of repeated checking on accuracy is small, we deem it unlikely that the large effects of repeated checking on memory confidence, vividness and detail were driven by the effect on accuracy (for a detailed explanation and discussion, see Section 4.3).

4.2. Virtual checking task 2.0

As explained in the introduction, earlier versions of the task contained a confound, where relevant checking was being carried out with the gas stove, but not with the light bulbs. Some studies lacked this confound (Dek, van den Hout, Giele, & Engelhard, 2010; Dek et al., 2014a, 2014b; Fowle & Boschen, 2011) yet still found the crucial interaction. The present study likewise was fully balanced, with relevant and irrelevant checking both being carried out with both sets of stimuli. As reflected by Fig. 3, there were strong and reliable drops in vividness, detail and confidence due to relevant checking, but no such drops occurred in the irrelevant checking control conditions. There were no effects of stimulus type (see Fig. 3), replicating earlier effects (Dek et al., 2014a, 2014b; Dek et al., 2010; Fowle & Boschen, 2011), indicating that the confound was immaterial to the effects previously observed and summarized in Fig. 1.

A further aim was to allow for assessing actual *behavior*, over and above self-report. The present paradigm does allow for this now. An unanticipated observation was the divergence between objective and subjective accuracy. Apparently, participants were fairly accurate at indicating which items they were asked to check (i.e., subjective accuracy). In fact, however, while carrying out the checking, participants' behavior was not always completely in correspondence with the actual instruction (i.e., objective accuracy). For instance, participants checked gas rings 1, 2, and 3, while the instruction was to check gas rings 1, 2, and 5. Subjective accuracy was around 85% on average, while objective accuracy was considerably lower (around 50%). We can only speculate as to why this was the case. Possibly, the fact that making errors had no negative consequences created some performance indifference and behavioral accuracy could increase if, for example, financial compensation for participating were to be made dependent on task performance.

Finally, we wanted the two conditions to be comparable in terms of effort it takes to complete the trials of the task, operationalized by response latencies. This proved extremely hard (see 3.1). Various pilots with small variations in the displays of the two tasks (light bulbs vs. gas rings) produced latency differences with one task being carried out faster or slower. Even the final task, reported here, did not completely reach this goal: it took participants (a little) more time to complete the light bulb trials than the gas ring trials (see Fig. 4); this was statistically significant. Had this, minor, in-equivalence difference had an effect on the subjective data (accuracy, confidence, vividness, and detail) there should have been stimulus effects in the ANOVAs presented under paragraph 3.3.1. No such effect was observed, showing that the slight latency-differences between the two tasks were irrelevant to the crucial outcome measures. Likewise, while the inequivalence, visible in Fig. 4, was reflected in a Condition main effect, the *slope* of the two lines in Fig. 4 are identical and there was no interaction effect between Condition and Time on the latency measure.

Although the aim of the study was not to test a particular hypothesis on OCD, the latency findings given in Fig. 4 warrant some discussion. The response latencies for both stimuli sped up quickly over the first few trials, approaching an asymptote after trials 10–15 and following a quadratic pattern. Apparently, then, the OCD-like perseveration as modelled in the paradigm, is subject to clear automatization, as suggested previously (van den Hout & Kindt, 2003a). Human information processing, including perception, memory, text comprehension and reading, elementary reasoning, motor behavior etc. is carried out with very little effort while the accuracy of these processes is typically taken for granted. Interestingly, the uncertainty that characterizes OCD patients typically relates to the trustworthiness of such automatic routines: Do I understand the text I am reading/hearing? The light seems off, but can I trust my eyes? Can I trust my hands not to strangle the baby? Can I trust the memory of me having closed the door? Importantly, the type of safety behavior employed by many OCD patients is some sort of repetition, like re-reading texts, staring at light switches or checking gas rings. Ironically, such repetition serves to *promote* the very automaticity that is distrusted (see Fig. 4) and, as indicated by Fig. 1b and 6, to reduce confidence in memory. Why OCD patients use repetition, provoking automatization, to combat distrust in automatic processes is a psychological puzzle requiring further study. The present paper provided some solutions and new insights, while generating challenges for new studies. We will end the discussion by summarizing and discussing these issues.

4.3. Solutions

The repeated checking paradigm previously suffered from some weaknesses: (1) there was a potential confound, (2) it was unclear whether the paradigm sub-tasks were equally demanding, and (3) the paradigm did not allow for assessing behavior. These problems were solved in the version of the task presented here,. The updated task is freely available online (see Supplementary Material) to facilitate use of the task for future research. One important avenue for future research is to use our updated task in a sample of participants who meet the criteria for OCD, to see whether similar automatization effects could be obtained.

4.4. New insights

From our findings, we obtained three major insights. The first new insight concerns the observed effect of memory accuracy in the meta-analysis. The majority of the studies included in the meta-analysis found strong effects on meta-memory (confidence, vividness and detail), but not on memory accuracy (only 3/17 found a significant (and medium) effect). However, when collapsed in the meta- analysis and with enhanced statistical power, it became clear that repeated checking does also reduce memory accuracy, with the effect being substantially smaller than the meta-memory effects. Could reductions in accuracy and in meta-memory have affected each other? A first option is that reductions in accuracy may have contributed to reductions in memory confidence, vividness and detail. With regard to memory *confidence*, this may be a plausible explanation for part of the effect. In particular, when filling out the forced choice accuracy data, individuals may have sensed that they were making errors, resulting in a diminished level of confidence in their accuracy ratings. However, it is less clear how reduced accuracy may have contributed to reduced *vividness and detail* as well. Moreover, the errors, as reported in the studies included in the meta- analysis, were made by only a minority of participants, making it unlikely that these individuals were responsible for the robust meta-memory effects. A second, alternative option is that reduced vividness, detail and confidence could have an profound effect on accuracy. If this were true, one would likewise expect robust reductions in accuracy, which were not observed in our results. In our view, the observed findings can be more convincingly be explained by repetition-induced automatization of behavior. Note that repetition-induced automatization of behavior reduces the ability to explicitly memory the nature of the automated behavior (and or its consequences; Schacter, 1987) and the reduced confidence found in the meta-analysis seems to echo this effect. Interestingly, the performance on forced choice memory tasks is much more resistant to the effect of automatization. As such, implicit memory may remain intact even when explicit memory is reduced or lost (Schacter, 1987). When repetition and automatization increases, eventually accuracy may suffer as well. Put differently, automatization may initially have a negative impact on explicit memory only, reflected in reduced meta-memory ratings, and later on implicit memory as seen in the subtle effects on the forced choice assessment of accuracy.

The second new insight concerns the observed discrepancy between objective and subjective memory accuracy. While participants remained relatively good in indicating which objects they were instructed to check (subjective accuracy), they were in fact less accurate in checking the right stimuli (objective accuracy), see Table 1.

Third, our findings confirm the van den Hout & Kindt (2003a) hypothesis that repeated checking renders the behavior an automatic routine (see slopes of the lines given in Fig. 4). Importantly, our study expands an earlier study by Dek et al., (2014a, 2014b), by demonstrating this automatization in a more direct way.

4.5. Challenges

Salkovskis (1998) notes that “OCD patients attempt to monitor closely and take control over processes that would otherwise operate in automatic and well-practiced ways”. Echoing this clinical observation, Soref, Dar, Argov, and Meiran (2008) found in an experimental study that individuals with OCD traits are less likely to shift from focused processing to parallel processing. This suggests an OCD-related distrust in automatic routines. The origin of this distrust is largely obscure, however the experience of normal automatic routines being untrustworthy can be extremely alarming. An OCD patient/psychiatrist thoughtfully labelled the experience an’emotional illusion’ (Oosterhoff, 2017, personal communication) likening the latter to the optical illusion of lines not running in parallel while in fact they do (see Fig. 5). One knows the lines are parallel, but this does not change the experience of the lines sloping down. Likewise, the OC experience of distrusting one's hands in terms of strangling the baby may feel as convincing as seeing the lines sloping down. While the latter illusion may be irrelevant to ones well- being, the emotional illusion is relevant. Not trusting the automatic routines, the patient may then try and repeat the behavior in a controlled, effortful way: re-reading with utmost concentration, now while speaking out loud, staring at the doorknob, trying to increase certainty by telling oneself: “it is really closed, I am seeing it is closed” and by adding tactile experience (touching the knob) to the visual and auditory information. If the patient were then to ask himself whether it is *credible* that the door is still open, he will kill the baby etc., the answer would be negative. The overwhelming uncertainty motivates patients to ask another question that seems related, but that is fundamentally different: “Is it, ultimately, still *possible* that the door is not closed or that I will strangle the baby?” In the final analysis, the answer here is affirmative. This may provide motivation to further increase certainty by repeating the act, if possible with even more concentration and effort. This leads to further research questions: under what condition are OCD patients uncertain about the effects of automatic routines? Would the distrust of automatic routines be a special case of an attenuated access to internal states (Dar, Lazarov, & Liberman, 2016)? We suggested that, in the realm of their worries, OCD patients replace the question about probability/credibility by questions about absolute certainty. This is a clinical impression. Would it survive

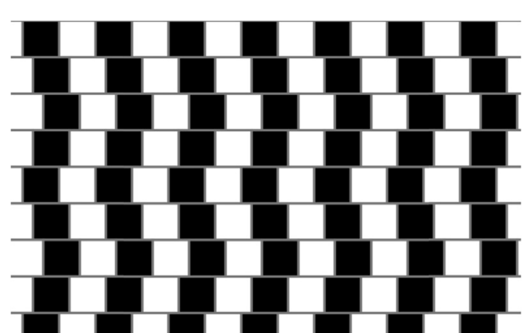


Fig. 5. The horizontal lines do not seem to be parallel (Café wall illusion).

critical tests? Questions about 'ultimate certainty' seem intrinsically uninformative, because the answer will always be negative. Would helping patients replace this uninformative question by the very informative question about credibility stimulate progress in cognitive behavioral therapy (CBT)? Would accepting the 'emotional illusion' as a curious phenomenon, and training oneself in not responding to it reduce its impact? There is room for fresh research.

Acknowledgements

The authors thank Roy van Kooten for the development of the improved experimental paradigm in Matlab, and Professor Ton de Jong for sharing the stimulus material from the ZAP project (http://zap.psy.utwente.nl/).

Appendix A

Search strategies per database

<u>PsycInfo: 111 hits</u>

1. memory/
2. memory.ab,ti."
3. check*".ab,ti."
4. repeat*".ab,ti.
5. 3 and 4
6. "persever*".ab,ti.
7. 3 and 6
8. 5 or 7
9. 1 or 2
10. 8 and 9

<u>PubMed: 113 hits</u>

(Memory [Title/Abstract] OR "Memory" [MeSH]) AND ((Check* [Title/Abstract] AND Repeat* [Title/Abstract]) OR (Check* [Title/Abstract] AND Persever* [Title/Abstract]))

<u>Embase: 214 hits</u>

#1'memory'/exp OR 'memory': ab,ti
#2'check*': ab,ti AND 'repeat*': ab,ti
#3'check*': ab,ti AND 'persever*': ab,ti
#4#2 OR #3 #5#1 AND #4

<u>OpenGrey: 52 hits</u>

check* AND (repeat* OR persever*)

Appendix B

See Appendix Table B1

Table B1
Explorative subgroup analyses on effects of repeated checking on confidence, vividness and detail.

		Confidence			Vividness			Detail		
		k	*g*	I^2	*k*	*g*	I^2	*k*	*g*	I^2
Overall		28	.887	58.29	27	.902	57.89	27	.872	54.72
Design										
	Mixed	23	.836	40.33	22	.862	43.80	22	.824	18.58
	Within	5	1.082	83.35	5	1.039	81.65	5	1.019	86.34
Participants										
	Healthy	25	.894	62.21	24	.940	60.46	24	.887	59.17
	OCD	3	.817	0	3	.526	0	3	.725	0
Trials										
	20	21	.980	62.21	20	.987	62.44	20	.949	54.29
	Other	7	.625	0	7	.679	0	7	.655	43.25
Stimuli										
	Gas vs. lights	14	.989	71.5	13	1.089	60.02	13	.974	57.20
	Other	14	.778	17.17	14	.726	40.72	14	.774	49.90
Checking										
	Virtual	21	.919	61.57	20	.972	57.52	20	.915	51.69
	Other	7	.771	43.02	7	.669	42.05	7	.749	58.39
Balancing										
	Partial	20	.929	65.79	19	1.005	56.70	19	.951	58.45
	Full	8	.822	20.25	8	.693	51.21	8	.718	34.66

Note: k = number of studies; g = Hedges' g; I^2 = index of heterogeneity.

Appendix C

See appendix Fig. C1

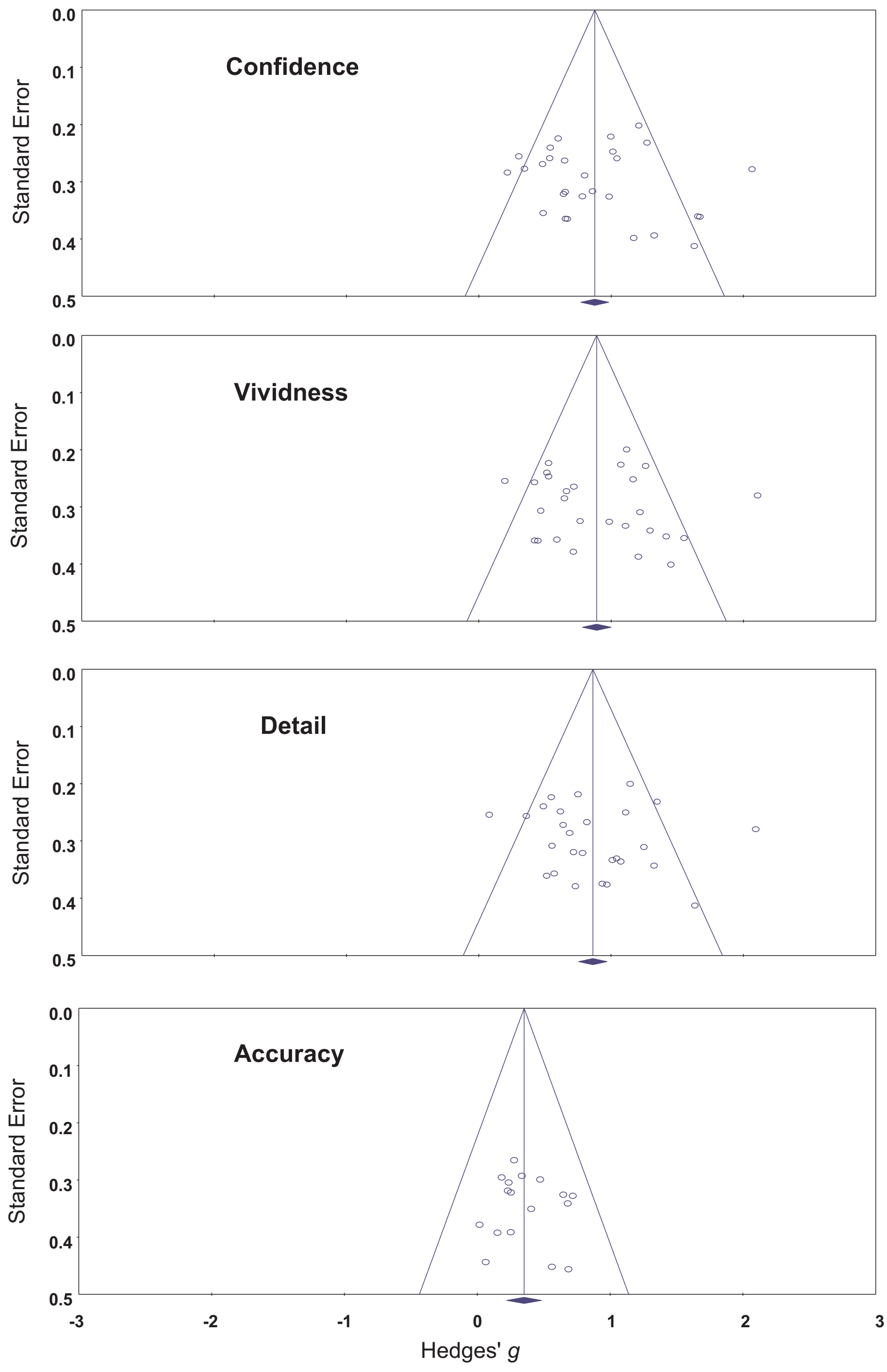


Fig. C1. Funnel plots of Standard Error by Hedges' *g* for confidence, vividness, detail and accuracy.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jocrd.2017.11.006.

References[1]

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* John Wiley & Sons, Ltd.http://dx.doi.org/10.1002/9780470743386.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Research Synthesis. *Methods, 8*(1), 5–18. http://dx.doi.org/10.1002/jrsm.1230.

(*)Boschen, M. J., & Vuksanovic, D. (2007). Deteriorating memory confidence, responsibility perceptions and repeated checking: Comparisons in OCD and control samples. *Behaviour Research and Therapy, 45*(9), 2098–2109. http://dx.doi.org/10.1016/j.brat.2007.03.009.

(*)Boschen, M. J., Wilson, K. L., & Farrell, L. J. (2011). Attenuating memory distrust in a repeated checking task. *Behaviour Research and Therapy, 49*(8), 466–471. http://dx.doi.org/10.1016/j.brat.2011.05.001.

Comprehensive meta-analysis (Version Version 3.3.070). Englewood, NJ: Biostat.

Cuijpers, P., Weitz, E., Cristea, I. A., & Twisk, J. (2016). Pre-post effect sizes should be avoided in meta-analyses. *Epidemiology and Psychiatric Sciences, 26*(4), 364–368. http://dx.doi.org/10.1017/S2045796016000809.

Dar, R., Lazarov, A., & Liberman, N. (2016). How can I know what I'm feeling? Obsessive-compulsive tendencies and induced doubt are related to reduced access to emotional states. *Journal of Behavior Therapy and Experimental Psychiatry, 52*, 128–137. http://dx.doi.org/10.1016/j.jbtep.2016.04.004.

Dar, R., Rish, S., Hermesh, H., Taub, M., & Fux, M. (2000). Realism of confidence in obsessive-compulsive checkers. *Journal of Abnormal Psychology, 109*(4), 673. http://dx.doi.org/10.1037/0021-843X.109.4.673.

(*)Dek, E. C., van den Hout, M. A., Engelhard, I. M., Giele, C. L., & Cath, D. C. (2015). Perseveration causes automatization of checking behavior in obsessive-compulsive disorder. *Behaviour Research and therapy, 71*, 1–9. http://dx.doi.org/10.1016/j.brat.2015.05.005.

(*)Dek, E. C., van den Hout, M. A., Giele, C. L., & Engelhard, I. M. (2010). Repeated checking causes distrust in memory but not in attention and perception. *Behaviour Research and Therapy, 48*(7), 580–587. http://dx.doi.org/10.1016/j.brat.2010.03.009.

(*)Dek, E. C., van den Hout, M. A., Giele, C. L., & Engelhard, I. M. (2014a). Automatization and familiarity in repeated checking. *Journal of Obsessive-Compulsive and Related Disorders, 3*(4), 303–310. http://dx.doi.org/10.1016/j.jocrd.2014.07.001.

(*)Dek, E. C. P., van den Hout, M. A., Giele, C. L., & Engelhard, I. M. (2014b). Automatization and familiarity in repeated checking: A replication. *Journal of Experimental Psychopathology, 6*(2), 149–167. http://dx.doi.org/10.5127/jep.045314.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. http://dx.doi.org/10.1111/j.0006-341x.2000.00455.x.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629.

(*)Fowle, H. J., & Boschen, M. J. (2011). The impact of compulsive cleaning on confidence in memory and cleanliness. *Journal of Anxiety Disorders, 25*(2), 237–243. http://dx.doi.org/10.1016/j.janxdis.2010.09.008.

*Giele, C. L., Engelhard, I. M., van den Hout, M. A., Dek, E. C. P., & Damstra, M. F. (2014). Repeated checking induces uncertainty about future threat. Practice does not make perfect: the ironic effects of compulsive perseveration. Retrieved from 〈https://dspace.library.uu.nl/handle/1874/294804〉.

Hermans, D., Engelen, U., Grouwels, L., Joos, E., Lemmens, J., & Pieters, G. (2008). Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behaviour Research and Therapy, 46*(1), 98–113. http://dx.doi.org/10.1016/j.brat.2007.11.001.

Hermans, D., Martens, K., De Cort, K., Pieters, G., & Eelen, P. (2003). Reality monitoring and metacognitive beliefs related to cognitive confidence in obsessive–compulsive disorder. *Behaviour Research and Therapy, 41*(4), 383–401. http://dx.doi.org/10.1016/S0005-7967(02)00015-3.

Higgins, J. P. T. & Green, S. (2011). Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration.

(*)Linkovski, O., Kalanthroff, E., Henik, A., & Anholt, G. (2013). Did I turn off the stove? Good inhibitory control can protect from influences of repeated checking. *Journal of Behavior Therapy and Experimental Psychiatry, 44*(1), 30–36. http://dx.doi.org/10.1016/j.jbtep.2012.07.002.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Macdonald, P. A., Antony, M. M., Macleod, C. M., & Richter, M. A. (1997). Memory and confidence in memory judgments among individuals with obsessive compulsive disorder and non-clinical controls. *Behaviour Research and Therapy, 35*(6), 497–505. http://dx.doi.org/10.1016/S0005-7967(97)00013-2.

(*)Medway, M., & Jones, M. K. (2013). Further evidence that repeated checking leads to reduced memory confidence, vividness and detail: New evidence that repeated object exposure also results in memory distrust. *Behaviour Change, 30*(03), 159–179. http://dx.doi.org/10.1017/bec.2013.16.

Oosterhoff, M. (2017) Personal communication.

Rachman, S. (2002). A cognitive theory of compulsive checking. *Behaviour Research and Therapy, 40*(6), 625–639. http://dx.doi.org/10.1016/S0005-7967(01)00028-6.

(*)Radomsky, A. S., Gilchrist, P. T., & Dussault, D. (2006). Repeated checking does cause memory distrust. *Behavior Research and Therapy, 44*, 306–316. http://dx.doi.org/10.1016/j.brat.2005.02.005.

(*)Radomsky, A. S., Dugas, M. J., Alcolado, G. M., & Lavoie, S. L. (2014). When more is less: Doubt, repetition, memory, metamemory, and compulsive checking in OCD. *Behaviour Research and therapy, 59*, 30–39. http://dx.doi.org/10.1016/j.brat.2014.05.008.

Ruscio, A. M., Stein, D. J., Chiu, W. T., & Kessler, R. C. (2010). The epidemiology of obsessive-compulsive disorder in the national comorbidity survey replication. *Molecular Psychiatry, 15*(*1*), 53. http://dx.doi.org/10.1038/mp.2008.94.

Salkovskis, P. M. (1998). Psychological approaches to the understanding of obsessional problems. In R. P. Swinson, M. M. Antony, S. Rachman, & M. A. Richter (Eds.). *Obsessive–compulsive disorder* (pp. 40). New York: Guilford Press.

Salkovskis, P. M., & Forrester, E. (2002). Responsibility. In O. Frost, & G. Steketee (Eds.). *Cognitive approaches to Obsessions and compulsions: theory, assessment and treatment* (pp. 45–63). Oxford: Pergamon Press.

Soref, A., Dar, R., Argov, G., & Meiran, N. (2008). Obsessive–compulsive tendencies are associated with a focused information processing strategy. *Behaviour Research and Therapy, 46*(12), 1295–1299. http://dx.doi.org/10.1016/j.brat.2008.09.001.

Tallis, F. (1995). *Obsessive compulsive disorder. A cognitive and neuropsychological perspective.* Chichester: Wiley.

(*)Toffolo, M. B., van den Hout, M. A., Radomsky, A. S., & Engelhard, I. M. (2016). Check, check, double check: Investigating memory deterioration within multiple sessions of repeated checking. *Journal of Behavior Therapy and Experimental Psychiatry, 53*, 59–67. http://dx.doi.org/10.1016/j.jbtep.2015.09.001.

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive–compulsive disorder. *Behaviour Research and Therapy, 39*(8), 913–927. http://dx.doi.org/10.1016/S0005-7967(00)00064-4.

van den Hout, M. A., Engelhard, I. M., & McNally, R. J. (2017). Thoughts on experimental psychopathology. *Psychopathology Review, 4*(2), 141–154. http://dx.doi.org/10.5127/pr.045115.

(*)van den Hout, M. A., & Kindt, M. (2003aa). Repeated checking causes memory distrust. *Behaviour Research and Therapy, 41*(3), 301–316. http://dx.doi.org/10.1016/S0005-7967(02)00012-8.

(*)van den Hout, M. A., & Kindt, M. (2003bb). Phenomenological validity of an OCD-memory model and the remember/know distinction. *Behaviour Research and Therapy, 41*(3), 369–378. http://dx.doi.org/10.1016/S0005-7967(02)00097-9.

(*)van den Hout, M., & Kindt, M. (2004). Obsessive–compulsive disorder and the paradoxical effects of perseverative behaviour on experienced uncertainty. *Journal of Behavior Therapy and Experimental Psychiatry, 35*(2), 165–181. http://dx.doi.org/10.1016/j.jbtep.2004.04.007.

*van den Woudenberg (2016). *Does repeated checking really cause memory distrust? Improving the computer task*. Unpublished manuscript.

(*)van Dis, E. A. M., & van den Hout, M. A. (2016). Not just right experiences as ironic result of perseverative checking. *Clinical Neuropsychiatry, 13*(6), 100–107.

[1] References with asterisks indicate studies included in the meta-analysis.