Published in final edited form as:
Cortex. 2020 August 01; 129: 510–525. doi:10.1016/j.cortex.2020.03.031.

Preventing the return of fear in humans using reconsolidation update mechanisms: A verification report of Schiller et al. (2010)

Anastasia Chalkia[a,b], **Lukas Van Oudenhove**[b,c], **Tom Beckers**[a,b,*]

[a]Centre for the Psychology of Learning and Experimental Psychopathology, Faculty of Psychology & Educational Sciences, KU Leuven, Tiensestraat 102 – box 3712, 3000 Leuven, Belgium

[b]Leuven Brain Institute, KU Leuven, Leuven, Belgium

[c]Laboratory for Brain-Gut Axis Studies (LaBGAS), Translational Research Centre for Gastrointestinal Disorders (TARGID), Department of Chronic Diseases, Metabolism, and Ageing, KU Leuven, O&N I Herestraat 49 – box 701, 3000 Leuven, Belgium

Abstract

In a highly influential report, Schiller and colleagues (2010) demonstrated long-lasting fear reduction in humans when conducting extinction training shortly following fear memory reactivation. While trying to experimentally replicate the critical conditions of Schiller et al. (2010, Experiment 1), we discovered several irregularities in their paper. Criteria for participant exclusion and the number of excluded participants were misreported; qualitative experimenter decisions actually determined their participant inclusions. Moreover, their statistical analyses were internally inconsistent. After corresponding with the original authors, we received their original data files, allowing us to replicate the reported analyses to verify their results. Here, we report the results of seven separate sets of analyses, three replicating the analyses reported by Schiller et al. (2010) and four applying more principled approaches to participant exclusion, thus including different subsets of the total datasets available, to deduce the influence of specific exclusions and experimenter decisions on the results. For Experiment 1, we were mostly able to replicate the analyses contained in the original report when applying the same qualitative exclusions. However, we found that all of the differences in fear recovery between reactivation-extinction and regular extinction reported by Schiller et al. (2010) were dependent on the qualitative exclusions that they made. With any of the principled approaches to participant exclusion, the degree of fear recovery was highly similar between groups. For Experiment 2, a similar analysis was not possible due to a lack of available data for the excluded participants. Hence, we conducted a verification analysis on the original sample only, which failed to confirm the differences in fear recovery reported by Schiller et al. (2010). Together with the re-analyses, we report a number of additional issues with the way Schiller et al. (2010) processed, analyzed, and reported their data that indicate that their results are unreliable and flawed.

[*]Address for correspondence: Tom Beckers, Faculty of Psychology & Educational Sciences, KU Leuven, Tiensestraat 102 – box 3712, 3000 Leuven, Belgium, Tel +32 16 32 61 34, Fax +32 16 32 60 99, tom.beckers@kuleuven.be.

Declarations of interest: none

Data availability
The datasets generated for this report are freely available on the OSF (https://osf.io/ayjqr/).

Keywords

reactivation-extinction; post-retrieval extinction; reconsolidation; memory updating

1 Introduction

In 2010, *Nature* published a paper that immediately attracted the attention of memory researchers and clinicians alike, as it introduced a non-invasive procedure that appeared to durably update maladaptive emotional memories and thereby permanently prevent the return of fear memory expression in humans (Schiller et al., 2010). The procedure exploited the notion of memory reconsolidation (Przybyslawski & Sara, 1997; Sara, 2000), a putative mechanism that allows for the restabilization of memory representations after their retrieval-induced destabilization. The proposed function of the destabilization-reconsolidation cycle is to allow the maintenance of memory relevance, by providing an opportunity for the updating of previously consolidated representations in light of new information (Lee, 2009). Building on this notion, Schiller and colleagues (2010) hypothesized that conducting extinction training while a conditioned fear memory was undergoing reconsolidation would lead to the updating of the original excitatory fear memory, instead of the creation of a separate inhibitory memory, which regular extinction training is postulated to do (Bouton, 2004). As a result, after reactivation-extinction an updated memory in which the excitatory CS-US memory is effectively replaced by an inhibitory extinction or CS-noUS memory would be formed. This updated memory would in turn counter the possibility of spontaneous recovery, renewal, or reinstatement of fear memory expression, phenomena that are typically observed after regular extinction training, and are thought to reflect the intact excitatory fear memory regaining dominance over the inhibitory extinction memory (Bouton, 2002).

In support of this idea, research had previously demonstrated a benefit of reactivation-extinction over regular extinction in preventing return of fear in rats (Monfils, Cowansage, Klann, & LeDoux, 2009). Translating that animal work, Schiller et al. (2010, Experiment 1)likewise reported an absence of recovery of conditioned fear responding in humans in a group that received memory reactivation followed 10 minutes later by extinction. On the contrary, fear memory expression recovered in a regular extinction group and a group that received extinction 6 hours following memory reactivation (by which time the memory is supposed to no longer be malleable). One year later, available participants were invited for a follow-up (1-year FU) test where they were exposed to unsignalled shocks to reinstate the conditioned fear memory, before being tested for fear responding (see Figure 1 for a schematic overview of Experiment 1). Those that had, 1 year earlier, received extinction 10 min after reactivation still demonstrated a persistent absence of fear memory expression. In a within-subjects variant of the same procedure, Schiller et al. (2010, Experiment 2) paired two different stimuli with shock and one day later reactivated only one of them before conducting extinction training for both. A further day later, participants exhibited recovery of fear responding following unsignalled shock presentations only for the stimulus that had not been reactivated prior to extinction, indicating that the reactivation-extinction effect was specific to the reactivated memory trace only (see Figure 2 for a schematic overview of Experiment 2).

The findings of Schiller et al. (2010) had massive impact, with their report accumulating over 600 citations according to Web of Science (and over 1,000 according to Google Scholar) by September 2019. In its wake, a multitude of conceptual replication studies have been published, some successful and others not, which illustrates the considerable time and resources that researchers have been investing in building on this effect and its translational potential over the past decade (for reviews see Auber, Tedesco, Jones, Monfils, & Chiamulera, 2013; Monfils & Holmes, 2018; Zuccolo & Hunziker, 2019; for a meta-analysis see Kredlow, Unger, & Otto, 2016).

In light of the impact of the results of Schiller et al. (2010) and their important basic and translational implications, we embarked on an independent, direct replication of the critical groups of Schiller et al. (2010, Experiment 1), which had not been done before (Chalkia et al., submitted). When we started collecting data, we found that we had to exclude close to 75% of our participants when applying the exclusion criteria described in Schiller et al. (2010), which was surprising given the very low rate of exclusions reported by these authors (8%). Ready to abandon our replication attempt, considering the effort it would require to complete a three-day fear conditioning study where 75% of participants had to be excluded, we requested the original data from Schiller et al. (2010) to try to figure out the cause of this unlikely difference in exclusion rates. Upon receiving the data (part of which have since been made available by the authors on the Open Science Framework (OSF), https://osf.io/jhu5c/), we found that the exclusion criteria reported by Schiller et al. (2010) had not been applied as stated; applying them to the original dataset similarly led to about 74% of exclusions for their Experiment 1. When we brought this to the attention of the authors, we were presented with a new set of exclusion criteria (Daniela Schiller & Elizabeth Phelps, personal communication, September 21, 2017; see Methods) that were noticeably different from the original ones. The new set of criteria were substantially more lenient and allowed for the inclusion of participants who showed no acquisition of differential conditioned responding (for an example, see Figure 3). Furthermore, whereas Schiller et al. (2010, Experiment 1) reported only 6 exclusions, in reality a much higher number of participants seemed to have been excluded, at different timepoints in the study. Finally, some data points contained in the dataset did not seem veridical, as the dataset contained occasional SCR responses to the shock US that seem physiologically implausible (including amplitudes of up to 625 μS) under the specifications of the Biopac SCR system used by Schiller et al. (2010) – typical values for SCR amplitude range from threshold to a maximum of around 8 μS in highly aversive procedures (Braithwaite, Watson, Jones, & Rowe, 2015).

One year after we brought the misreporting of exclusion criteria and number of exclusions to the authors' attention, an addendum was published in *Nature* that aimed to clarify these issues (Schiller et al., 2018). However, this addendum spawned further uncertainty about the way data were handled in the original study. First, whereas Schiller et al. (2010) initially reported having recruited 71 participants in total for Experiment 1 (6 of which were excluded), in the addendum they disclosed to having recruited 126 participants, 61 of which were excluded at different stages of data collection. Yet, as they proceeded to list the specific reasons for the different exclusions, the exclusions seem to add up to 66 rather than 61. Further, the inclusion criteria they reported in the addendum were slightly different again from the criteria that they had confirmed to us in 2017. Moreover, the authors stated that "exclusion or inclusion of participants during the study was based on a *qualitative* assessment of trial-by-trial SCR as reflecting a pattern consistent with successful acquisition and extinction" and that the criteria listed were merely meant to provide a "*characterization* of the final dataset, corresponding to the *qualitative evaluations* made in the process of data collection and analysis" (italics added).

The most interesting point conveyed by the addendum and its online supplement is indeed the observation that inclusion criteria were not applied rigorously. At several occasions, the decision was made to include participants who did not fit any of the criteria; at the same time, other participants were excluded despite fitting the criteria. The reasons provided for those decisions raise further questions. For example, 3 participants who met at least one of the acquisition criteria for Experiment 1 were reportedly excluded "for [displaying] idiosyncratic responses not observed in other participants in the sample (e.g., conditioned response on last acquisition trial was negative or almost equal to zero)" (Schiller et al., 2018, p. 2). Yet when inspecting the final dataset, one can observe that 8 included participants also displayed a negative conditioned response on the last acquisition trial, and another 11 included participants exhibited responding equal or almost equal to zero. It is unclear why those participants were then not excluded for the same reason of idiosyncratic responding.

Given that the discrepancies in the original manuscript and its 2018 addendum and visual inspection of the data raise so many questions, independent verification of the results reported by Schiller et al. (2010) seemed warranted. As indicated above, the original datasets of Schiller et al. (2010) were partly made available on the OSF, containing the data from 65 participants who were included for analysis in Experiment 1 as well as those of 19 participants who were excluded[1] but for whom complete 3-day data were available. The data for the 1-year follow-up to Experiment 1 were not included on the OSF, but they were sent to us by the last author (Elizabeth Phelps, personal communication, July 23, 2017). For Experiment 2, only the data of 18 included participants were available online. The data files sent to us by the authors did not contain the data of all 52 participants who were reportedly excluded during data collection but did contain the data of 6 participants who had been excluded during (4) or after (2) data collection.

We first set out to replicate all the analyses reported by Schiller et al. (2010) exactly, in order to verify their observations. In addition, we conducted 3 more sets of analyses for Experiment 1 and 1 for the 1-year follow-up to Experiment 1, each using a different rigorous set of participant exclusion criteria, in order to evaluate the influence of specific exclusions on the overall strength of the reactivation-extinction effect. With these verification analyses, we aim to provide an assessment of the accuracy and robustness of the results reported by Schiller et al. (2010) and of how experimenter decisions regarding participant exclusions, data processing and statistical choices may have shaped them.

[1] The full data of the 19 excluded participants were uploaded to the OSF, but their respective group allocations were not reported. However, this information was available in the data files we received from the original authors (Elizabeth Phelps, personal communication, July 23, 2017).

2 Methods

2.1 Processing of SCR data

The dataset of Schiller et al. (2010, Experiment 1 – Spontaneous recovery) was downloaded from the OSF. Raw SCR data were already pre-processed by the original authors, and the OSF dataset contained square-root transformed SCRs for CS+ and CS- trials separately. We proceeded to process the data in exactly the same way as reported by Schiller et al. (2010). We first standardized the data by dividing the square-root transformed value for each trial by that participant's mean square-root transformed US response (averaging over all US responses during acquisition). Only unreinforced acquisition trials were included, to exclude any influence of shock-elicited responding. Additionally, per the first author's instructions (Daniela Schiller, personal communication, September 13, 2017), the first CS- extinction trial was removed from the analysis, as were the first CS+ extinction trial in the no-reminder group and the CS+ reactivation trial in the 10 min and 6 h groups, resulting in a total of 10 CS+/CS- extinction trial pairs for all groups. Finally, the first CS- re-extinction trial was also removed, to prevent the influence of orienting responses. A differential score was then calculated for each trial pair, by subtracting CS- amplitudes from CS+ amplitudes for corresponding trials. The final dataset thus included 10 differential responses for each phase (i.e., acquisition, extinction, re-extinction).

The dataset of Schiller et al. (2010, Experiment 1 – Reinstatement; 1-year FU) is not available online. Those data were however contained in the files sent to us by the last author (Elizabeth Phelps, personal communication, July 23, 2017), and were already standardized by the original authors. Following the reported analysis pipeline (Schiller et al., 2010), a differential score (CS+ minus CS-) was computed for the end of re-extinction (trials 9 and 10), and after removal of the first CS+ and first CS- trials of reinstatement, the subsequent 4 trials were averaged in a differential score of post-reinstatement responding. We calculated an index of reinstatement by subtracting the differential score for the end of re-extinction from that of post-reinstatement responding. The data of the 6 h and no-reminder groups were then collapsed together.

The dataset of Schiller et al. (2010, Experiment 2) is available alongside that of Experiment 1 in the OSF repository. Data were already standardized and required no further processing.

2.2 Statistical Analyses

2.2.1 Experiment 1 – Spontaneous recovery—

Four different statistical analyses were performed on the processed SCR data, each using different criteria for participant inclusion (see 2.3). Otherwise, all analyses copied the analysis pipeline reported by Schiller et al. (2010). Differential scores were averaged across the first and last half (first 5 trials, last 5 trials) of each phase (acquisition and extinction) and subjected to a repeated measures analysis of variance (rm-ANOVA) with Group (10 min, 6 h, no-reminder) as a between-subjects factor and Time (first half, last half) as a within-subjects factor. To assess the decrease of fear responding from acquisition to extinction, an rm-ANOVA was used with Group as a between-subjects factor and Time (last half of acquisition, last trial of extinction) as a within-subjects factor. Spontaneous recovery was assessed with an rm-ANOVA using

Group as a between-subjects factor and Time (last trial of extinction, first trial of re-extinction) as a within-subjects factor[2]. Follow-up paired-sample *t* tests were conducted for each group separately to assess successful acquisition (comparing CS+ to CS- responding during the last half of acquisition), extinction (comparing CS+ to CS- responding during the last trial of extinction), the decrease in fear responding (comparing differential responding on the last half of acquisition to the last trial of extinction), and spontaneous recovery (comparing differential responding on the last trial of extinction to the first trial of re-extinction)[3]. Note that in line with Schiller et al. (2010), differential scores were used for the follow-up *t* tests examining the decrease in fear responding and spontaneous recovery, yet for acquisition and extinction CS+ values were compared to CS- values (i.e., the differential scores were not used).

In addition to the frequentist analyses, Bayesian analyses were conducted for the critical spontaneous recovery comparison. The default values in JASP (version 0.9.2; JASP Team, 2019) were used to quantify the relative evidence that the data came from the null versus an alternative model. Specifically, inverted Bayes factors (BF_{10}) were used to evaluate main effects, by comparing a model that includes the factor Time to a null model without the factor Time, and a model that includes the factor Group to a null model without the factor Group. To quantify the relative evidence for the presence of an interaction effect, compared to its absence, we compared a model which incorporates the interaction between Time and Group to a model that includes only the main effects (see 'Analyses of Effects' table in JASP). For these analyses, the Bayesian inclusion factor is reported ($BF_{\text{inclusion}}$). Jeffreys's evidence categories for interpretation (Jeffreys, 1961), with slightly modified categories as reported by Wetzels et al. (2011), were taken as the standard for evaluation of the reported Bayes Factors.

2.2.2 Experiment 1 – Reinstatement (1-year FU)—

To assess reinstatement, we performed an rm-ANOVA with Group (10 min, 6 h/no-reminder) as a between-subjects factor and Time (end of re-extinction, post-reinstatement) as a within-subjects factor. Follow-up (one-tailed) paired *t* tests compared the end of re-extinction to post-reinstatement responding in each group separately. Additionally, (one-tailed) independent samples *t* tests were used to compare post-reinstatement between groups and to evaluate between-group differences for the index of reinstatement. One-tailed tests were used here as this was the choice of the original authors for this part of the analyses only.

2.2.3 Experiment 2—

Prior to analyses, in line with the pipeline of Schiller et al. (2010) data for each stimulus (CSa+, CSb+, CS-) were averaged across the first and last half of the acquisition (trials 1-4, 5-8) and extinction (trials 1-5, 6-10) phases. For the re-extinction phase, the data were not used in full (total of 10 trials); instead, averages were made using the first 4 and next 4 trials (trials 1-4, 5-8). Rm-ANOVAs were conducted for each phase including within-subject factors Stimulus (CSa+, CSb+, CS-) and Time (first half, last half). Follow-up paired *t* tests were used to further assess acquisition (comparing CSa+ to CS-,

[2] This ANOVA deviates from what was reported in Schiller et al. (2010) but is in line with the follow-up analysis reported for the effect under investigation.

[3] This *t* test is consistent with what Schiller et al. (2010) reported and correctly follows-up the rm-ANOVA mentioned above.

CSb+ to CS-, CSa+ to CSb+; during the last half of the phase), extinction (comparing CSa+ to CS-, CSb+ to CS-, CSa+ to CSb+; on the last trial of the phase), and the reduction of fear (comparing the last half of acquisition to the last trial of extinction for all 3 stimuli). As with Experiment 1, the re-extinction ANOVA was not followed up as such; instead, recovery of fear was assessed through a paired *t* test comparing the last trial of extinction to the first trial of re-extinction for all 3 stimuli separately.

Greenhouse-Geisser corrections were applied in case of a violation of sphericity in the ANOVAs. An alpha level of .05 was set for all analyses, which were conducted using JASP (version 0.9.2; JASP Team, 2019). To make our results readily comparable to those of Schiller et al. (2010), *p* values below .05 are described as significant, while those above .05 are reported as non-significant.

2.3 Participant inclusions/exclusions

2.3.1 Experiment 1 – Spontaneous recovery analyses using original inclusions as reported in Schiller et al. (2010) (N = 65)—

In the original report (Schiller et al., 2010), 65 participants were included in the final analyses. The criteria for including these participants were specified in the supplement to the addendum (Schiller et al., 2018, pp. 1-3) as follows: "…exclusion or inclusion of participants during the study was based on qualitative judgements [made in the process of data collection and analysis] of trial-by-trial SCR as reflecting a pattern consistent with successful acquisition and extinction. Specifically, 'successful acquisition and extinction' for the participants in Study 1 can be translated to a chain of logical "IF" statements: A) If during acquisition the differential CS response (CS+ minus CS-) was below an individually standardized cut-off (value of 0.1 divided by the mean US response) on: (1) the first half of acquisition, (2) the second half acquisition, (3) the last trial of acquisition, and (4) the increase from the first to last trial of acquisition. When all of these criteria were met then acquisition was deemed as failed, otherwise the participant was included. B) If during extinction the differential CS response (CS+ minus CS-) was above the individually standardized cut-off on: 1) the second half of extinction, 2) the last trial of extinction, and 3) below cutoff in the decrease from the first to the last trials or halves of extinction. When all of these criteria were met then extinction was deemed as failed, otherwise the participant was included. C) In a limited number of cases, a decision was made to (a) include 2 participants that did not meet criteria for acquisition because the numerical distance from the cutoff was negligible (<0.005), or all measurable differential responses were positive or equal to zero, (b) include 1 participant that did not meet criteria for extinction because the numerical distance from the cut-off for a single criterion was negligible (.01), (c) exclude 3 participants despite not meeting all acquisition exclusion criteria for idiosyncratic responses not observed in other participants in the sample (e.g., conditioned response on last acquisition trial was negative or almost equal to zero) and (d) exclude 6 participants despite not meeting all extinction exclusion criteria, for idiosyncratic responses not observed in other participants in the sample (e.g., full reversal of the differential response, which might inflate the recovery index)." These qualitative decisions led to the exclusion of 19 participants for acquisition and extinction combined; when examining the final dataset, it is not always clear whether a particular participant was excluded for reasons of acquisition or extinction.

2.3.2 Experiment 1 – Spontaneous recovery analyses applying the inclusion criteria reported in Schiller et al. (2018) (N = 74)—As can be noted above, 8 separate criteria were reported in the addendum for considering acquisition and extinction successful, but those criteria were not applied rigorously. In our second set of analyses, we applied these criteria to the available dataset strictly as listed, without any qualitative decisions, which led to the exclusion of 10 participants (*n* = 6 for acquisition, *n* = 4 for extinction).

2.3.3 Experiment 1 – Spontaneous recovery analyses applying the initial inclusion criteria reported in Schiller et al. (2010) (N = 35)—The supplementary information of Schiller et al. (2010, p. 1) states that exclusions "…were based on the differential responses to the CS+ and CS- in the second half of acquisition and extinction. That is, subjects were excluded if during acquisition the difference was in the opposite direction (CS- > CS+) or smaller than 0.1 μS. Subjects were also excluded if during extinction the difference was in the opposite direction (CS+ > CS-) or larger than 0.1 μS." While from the 2018 addendum it is clear that these criteria were not used as such, for a third set of analyses we did actually apply these criteria to the dataset, which led to the exclusion of 49 participants (*n* = 21 for acquisition, *n* = 28 for extinction)[4].

2.3.4 Experiment 1 – Spontaneous recovery analyses on the full sample (N = 84)—Considering the shifting exclusion criteria reported by the authors and their failure to apply them rigorously, for a last set of planned analyses we included the full sample that was made available on the OSF.

2.3.5 Experiment 1 – Reinstatement analyses (1-year FU)— Schiller et al. (2010) reported having located 23 participants for their 1-year follow-up reinstatement test, of which 4 were excluded because of failure to re-extinguish after the spontaneous recovery experiment (differential score > 0.2 at the end of re-extinction) or due to SCR non-responding. In the files sent to us by the last author (Elizabeth Phelps, personal communication, July 23, 2017), we were able to locate data for 24 participants, of whom 1 could be excluded for non-responding, and 1 for a differential SCR > 0.2 at the end of re-extinction. Thus, we conducted one set of analyses on the remaining sample of *N* = 22. To further verify the reinstatement analysis (1-year FU) reported Schiller et al. (2010), which was carried out on 19 participants, we also set out to identify and exclude the 5 (rather than 4) participants who had been excluded by the original authors and conducted a second set of analyses on the resulting sample of *N* = 19 (see 3.5). Of the 5 excluded participants, 1 conformed to criteria for non-responding, but it is unclear why the other 4 were excluded, as none of them displayed a differential SCR > 0.2 at the end of re-extinction. Note that the 1 participant with a differential SCR > 0.2 at the end of re-extinction who we identified and excluded above was not actually excluded by Schiller et al. (2010).

[4] For this set of analyses, we did not apply the criterion for extinction which would exclude participants if responding was in the opposite direction (CS+ > CS-) on the last half of extinction. Examining the graphs of Schiller et al. (2010, Supplementary Information) and the corresponding *JoVE* report (Schiller, Raio, & Phelps, 2012), it was evident that this criterion was not used. If we were to apply it, 13 extra participants would have to be excluded for extinction, yielding a dataset of *N* = 22, which could not be reliably analyzed.

2.3.6 Experiment 2 analyses—In Schiller et al. (2010), the authors stated that they recruited 21 participants for Experiment 2, of whom 3 were excluded because they did not acquire a differential response to 1 of the 2 CS+s (thus, CS- > CSa+ or CS- > CSb+). Yet, in Schiller et al. (2018, p. 7), the total sample and exclusion criteria were updated, listing a total of 70 participants recruited, of which “52 were excluded at different stages of data collection for either being a non-responder or failing to acquire or extinguish conditioned responses [equally] to both CS+ stimuli.” Once again, “inclusion of participants was based on qualitative assessment of trial-by-trial SCR.” As the OSF dataset contained only the data of the included participants, we requested the data of the excluded participants from the corresponding author so that we could perform similar analyses as above. In personal communication (Elisabeth Phelps, December 14, 2019), she informed us that they had only retained data of participants who had full 3-day data, and that all exclusions for this study occurred after Day 1 or Day 2 of the study; hence, there were no further data available. Yet when examining the files sent to us before, we were able to locate data records for 6 additional participants (2 of which contained full 3-day data and 4 containing Day 1 acquisition data only; see Discussion). Given that we could not locate the data of all 52 excluded participants, we conducted analyses only on the 18 participants who were included in the original final sample.

3 Results

3.1 Experiment 1 – Spontaneous recovery analyses using original inclusions as reported in Schiller et al. (2010) (N = 65)

As reported in Schiller et al. (2010), these participants successfully acquired differential responding on the first day (main effect of time, $F(1, 62) = 9.92$, $p = .003$, $\eta_p{}^2 = 0.14$), with no significant differences between the groups (group x time, $F(2, 62) < 1$). Follow-up *t* tests confirmed that all groups responded more strongly to the CS+ than to the CS- in the last half of acquisition (see Table 1 for an overview of all *t* test analyses in this section). Differential responding extinguished over time (main effect of time, $F(1, 62) = 23.99$, $p < .001$, $\eta_p{}^2 = 0.28$),but unlike what was reported by Schiller et al. (2010), we observed a difference between the groups (group x time, $F(2, 62) = 4.96$, $p = .01$, $\eta_p{}^2 = 0.14$). In the first half of extinction, differential responding differed between the groups (main effect of group, $F(2, 62) = 3.35$, $p = .042$, $\eta_p{}^2 = 0.10$), while by the last half of extinction the groups were not significantly different (main effect of group, $F(2, 62) < 1$). Post-hoc comparisons with Bonferroni correction indicated that participants in the 6 h group exhibited less differentiation between CS+ and CS- at the beginning of extinction ($M = 0.07$, $SD = 0.10$) than those in the 10 min group ($M = 0.25$, $SD = 0.37$; $t(62) = 2.51$, $p = .044$, $d = 0.68$); no other comparisons were significant (see Figure 4). Further, and in line with Schiller et al. (2010), paired-sample *t* tests did not reveal significant differential responding by the last trial of extinction in any of the groups. The decrease in differential responding from the last half of acquisition to the last trial of extinction was significant across groups, with no significant between-group differences (main effect of time, $F(1, 62) = 38.63$, $p < .001$, $\eta_p{}^2 = 0.38$; group x time, $F(2, 62) < 1$). Follow-up *t* tests further confirmed a significant reduction in all the groups.

As a critical test for spontaneous recovery, Schiller et al. (2010, p. 50) reported "Spontaneous recovery was assessed using a two-way ANOVA with main effects of group (10 min, 6 h, and no-reminder) and time (early and late phase of re-extinction, defined by the mean first 4 responses versus the subsequent 4, respectively) revealing a significant main effect of time (F(1,62) = 6.26, $p < 0.05$), and a group × time interaction (F(2,62) = 4.63, $p <$ 0.05). Follow up t tests compared the differential responses between the last trial of extinction and the first trial of re-extinction." Schiller et al.'s choice of ANOVA model here is unusual, as it does not evaluate spontaneous recovery of differential conditioned responding from the end of extinction to the beginning of the test day (Day 3), but rather evaluates differences in the degree of re-extinction on the third day of the study, while using only part of the phase (8 out of 10 trials; see Figure 4 for trial-by-trial SCR data). It is then followed up with t tests that do not actually follow up on this ANOVA, as they do not compare the same timepoints. Regardless, we were able to exactly replicate this analysis (main effect of time, F(1, 62) = 6.27, p = .015, $\eta_p{}^2 = 0.09$; group x time, F(2, 62) = 4.63, p = .013, $\eta_p{}^2 = 0.13$). We then conducted the omnibus test that would have been the more obvious predecessor to the t tests reported by Schiller et al. (2010), comparing the last trial of extinction to the first trial of re-extinction, and found evidence for spontaneous recovery across groups (main effect of time, F(1, 62) = 13.20, $p < .001$, $\eta_p{}^2 = 0.18$), without significant between-group differences (group x time, F(2, 62) = 2.49, p = .091, $\eta_p{}^2 = 0.07$; see Figure 5A). Follow-up t tests corroborated the values reported by Schiller et al. (2010), showing that differential responding did not significantly change from the end of extinction to the beginning of re-extinction in the 10 min group, whereas it did increase significantly in the other 2 groups. Of note, while we were able to reproduce the t statistics for the 10 min and 6 h groups exactly to the second decimal place, the t statistic we obtained for the no reactivation group deviated slightly from the one reported by Schiller et al. (2010).

Bayesian analysis indicated that there was decisive evidence in support of a model including a change in SCR over time compared to the null model (BF_{10} = 180.32). Anecdotal evidence was found in favor of the null model including no differences in responding between the three groups versus an alternative model (BF_{10} = 0.37). Finally, it was about equally likely that the data came from a model containing an interaction term than from a model with only main effects ($BF_{\text{inclusion}}$ = 1.15). This implies that there is about equal evidence for or against there being differences between groups in spontaneous recovery in this sample.

All in all, we were able to almost exactly reproduce the statistical patterns and values reported by Schiller et al. (2010). This suggests that we processed the original data correctly. The small differences between the two sets of analyses do suggest that for some comparisons the authors used somewhat different variables than what was stated in their methods (see Discussion). Finally, it is noteworthy that Schiller et al. (2010) did not report an rm-ANOVA examining spontaneous recovery (which failed to yield a significant difference in spontaneous recovery between the groups) but instead reported a comparison of (partial) re-extinction (which did differ significantly between the groups).

3.2 Experiment 1 – Spontaneous recovery analyses applying the inclusion criteria reported in Schiller et al. (2018) (N = 74)

In our second set of analyses, results for acquisition and extinction were comparable to the first set of analyses. Participants displayed successful acquisition, without group differences (main effect of time, $F(1, 71) = 8.92$, $p = .004$, $\eta_p{}^2 = 0.11$; group x time, $F(2, 71) < 1$). In the last half of acquisition participants exhibited stronger responding to the CS+ than to the CS- in all groups (see Table 2 for an overview of all *t* test analyses in this section). Differential responding diminished from the first to the last half of extinction (main effect of time, $F(1, 71) = 21.73$, $p < .001$, $\eta_p{}^2 = 0.23$); a significant difference between the groups was again observed (group x time, $F(2, 71) = 4.93$, $p = .01$, $\eta_p{}^2 = 0.12$). Differential responding differed between the groups across the first half of extinction (main effect of group, $F(2, 71) = 3.21$, $p = .046$, $\eta_p{}^2 = 0.08$), but not during the second half of extinction (main effect of group, $F(2, 71) < 1$). Post-hoc Bonferroni corrected *t* tests were all non-significant. Again, by the last trial of extinction participants no longer differentiated significantly between CS+ and CS- in any of the groups. Differential responding decreased from the last half of acquisition to the last trial of extinction, without significant differences between groups (main effect of time, $F(1, 71) = 19.46$, $p < .001$, $\eta_p{}^2 = 0.22$; group x time, $F(2, 71) = 1.54$, $p = .22$, $\eta_p{}^2 = 0.04$). Follow-up *t* tests confirmed a reduction in differential responding in the 10 min group and in the no-reminder group, but not in the 6 h group.

Using this set of inclusion criteria, significant spontaneous recovery was not observed, as differential responding did not differ from the last trial of extinction to the first trial of re-extinction across groups (main effect of time, $F(1, 71) = 2.19$, $p = .14$, $\eta_p{}^2 = 0.03$; group x time, $F(2, 71) = 1.10$, $p = .34$, $\eta_p{}^2 = 0.03$; see Figure 5B). Follow-up *t* tests confirmed these findings, indicating no significant differences between the two timepoints in any of the groups. Bayesian analyses further corroborated these findings as there was anecdotal evidence in favor of the null model suggesting no change in SCR over time, compared to the alternative model ($BF_{10} = 0.80$). Anecdotal evidence also supported a model with no differences between the groups, versus an alternative model where groups differed ($BF_{10} = 0.37$). Lastly, a model containing the interaction term was only 0.34 times as probable as a model with only main effects, providing anecdotal evidence in support of the null model of no interaction between group and time ($BF_{\text{inclusion}} = 0.34$). Once again, compelling evidence favoring an interaction between group and time was not found. With these analyses, where the inclusion criteria reported in the addendum (Schiller et al., 2018) were rigorously applied, we conclude that spontaneous recovery of conditioned fear responding did not reliably differ between the groups.

3.3 Experiment 1 – Spontaneous recovery analyses applying the initial inclusion criteria reported in Schiller et al. (2010) (N = 35)

With this set of analyses, successful acquisition was again found on the first day (main effect of time, $F(1, 32) = 11.26$, $p = .002$, $\eta_p{}^2 = 0.26$), with no significant group differences (group x time, $F(2, 32) < 1$). However, examining the last half of acquisition for each group separately showed that the 10 min group did not exhibit significant CS+/CS- differentiation (see Table 3 for an overview of all *t* test analyses in this section). The other two groups demonstrated significantly higher fear responding to the CS+ than to the CS-. As the criteria for extinction were much stricter in this analysis compared to the previous ones, extinction was effective for all groups, with no significant group differences (main effect of time, $F(1, 32) = 23.96$, $p < .001$, $\eta_p^2 = 0.43$; group x time, $F(2, 32) = 1.86$, $p = .17$, $\eta_p^2 = 0.10$). By the last trial of extinction, CS+/CS- differentiation was no longer significant in any of the groups. Differential fear responses declined from the last half of acquisition to the last trial of extinction (main effect of time, $F(1, 32) = 18.71$, $p < .001$, $\eta_p^2 = 0.37$), without significant group differences (group x time, $F(2, 32) < 1$). Follow-up *t* tests for each group separately further corroborated this decline in the 6 h and no-reminder groups, but not in the 10 min group.

Analogous to the previous analysis (see 3.2), significant spontaneous recovery was not observed across groups, as differential responding was not reliably different between the last trial of extinction and the first trial of re-extinction (main effect of time, $F(1, 32) < 1$; group x time, $F(2, 32) < 1$; see Figure 5C). Follow-up *t* tests in each group were in the same line, as all three comparisons were non-significant. Bayesian analyses pointed to a similar pattern of results. Anecdotal evidence was found in favor of the null model compared to the alternative model, suggesting no differences in SCR over time ($BF_{10} = 0.54$). Substantial evidence was obtained for the null model versus an alternative model, indicating no differences between the groups ($BF_{10} = 0.18$). Data were only 0.48 times as likely to come from a model including an interaction between group and time, providing anecdotal evidence slightly favoring a null model of no interaction between group and time ($BF_{\text{inclusion}} = 0.48$). As such, using this set of criteria also leads to the conclusion that spontaneous recovery was effectively prevented in all of the groups.

3.4 Experiment 1 – Spontaneous recovery analyses on the full sample (N = 84)

When the entire sample available on OSF was included for analysis, differential acquisition was again observed across groups, without between-group differences (main effect of time, $F(1, 81) = 13.17$, $p < .001$, $\eta_p^2 = 0.14$; group x time, $F(2, 81) < 1$). Interestingly, using the full sample with no exclusion criteria for acquisition actually led to the strongest differential increase in responding across the acquisition phase. Further evidence was provided by the follow-up *t* tests showing stronger CS+ than CS- responding for all groups in the last half of acquisition (see Table 4 for an overview of all *t* test analyses in this section). Differential responding declined during extinction (main effect of time, $F(1, 81) = 16.03$, $p < .001$, $\eta_p^2 = 0.17$), but again, a difference between the groups was observed (group x time, $F(2, 81) = 3.79$, $p = .027$, $\eta_p^2 = 0.09$). Like in the previous analyses reported (see 3.1 and 3.2), this significant interaction was due to differences between the groups during the first half of extinction (main effect of group, $F(2, 81) = 4.23$, $p = .018$, $\eta_p^2 = 0.10$); by the second half responding no longer differed significantly between groups (main effect of group, $F(2, 81) < 1$). Follow-up Bonferroni corrected *t* tests showed a significant difference between the 10 min ($M = 0.32$, $SD = 0.52$) and 6 h groups ($M = 0.08$, $SD = 0.11$; $t(81) = 2.68$, $p = .027$, $d = 0.65$). Once more, by the last trial of extinction, none of the groups showed significant differentiation between CS+ and CS- anymore. A decrease in differential responding was observed across groups from late acquisition to the last trial of extinction, without between-group differences (main effect of time, $F(1, 81) = 26.64$, $p < .001$, $\eta_p^2 = 0.25$; group x time, $F(2, 81) = 1.49$, $p = .23$, $\eta_p^2 = 0.04$). Follow-up *t* tests confirmed this decline in the 10 min group and the no-reminder group, yet, the 6 h group did not exhibit a significant differential fear decrease.

With the full dataset, spontaneous recovery was observed across the groups, without between-group differences (main effect of time, $F(1, 81) = 4.42$, $p = .039$, $\eta_p^2 = 0.05$; group x time, $F(2, 81) < 1$; see Figure 5D). Follow-up *t* tests revealed that, although all groups showed a numerical increase in their differential responding from the last trial of extinction to the first trial of re-extinction, for none of the individual groups did this increase reach statistical significance. Bayesian analysis suggested substantial evidence in support of a model indicating a change in SCR over time versus the null model ($BF_{10} = 4.03$). Substantial evidence was found for the null model compared to the alternative model, suggesting no group differences ($BF_{10} = 0.11$). With regard to the interaction between group and time, it was only 0.31 times more likely that the data came from a model containing this interaction rather than a model without, demonstrating substantial evidence in support of a model with no interaction between group and time ($BF_{inclusion} = 0.31$).

Considering the significant increase over time in differential responding with no group effects or interactions, and the substantial evidence in the same direction provided by Bayesian analysis, using the full sample available, we obtained no evidence for a benefit of reactivation-extinction within the reconsolidation window, over regular extinction, or over reactivation-extinction outside of the reconsolidation window.

3.5 Experiment 1 – Reinstatement analyses (1-year FU)

The rm-ANOVA comparing the end of re-extinction to post-reinstatement between groups ($N = 22$) revealed no significant effects (main effect of time, $F(1, 20) = 2.16$, $p = .16$, $\eta_p^2 = 0.10$; main effect of group, $F(1, 20) < 1$; group x time, $F(1, 20) < 1$). Follow-up (one-tailed) *t* tests per group likewise failed to yield evidence for an increase in responding in either group (see Table 5 for an overview of all *t* test analyses in this section). (One-tailed) independent-samples *t* tests did not yield evidence for a between-group difference either, as there were no significant differences between groups in the reinstatement index or in post-reinstatement SCR responding. Taken together, these results suggest that the two groups were not differentially sensitive to the reinstatement manipulation (see Figure 6).

Schiller et al. (2010) did report significant between-group differences, after participant exclusions that left a sample size of $N = 19$. We were able to locate the excluded participants in the files sent to us, but even in that sample, we were unable to replicate their statistical results. Closer scrutiny of the original data files pointed to incongruities in the processing of the original data and their reporting. Unlike what was stated in the supplement to Schiller et al. (2018), it were not the first CS+ and CS- trials of reinstatement testing that were discarded; instead the first CS+ trial was excluded for some participants (with trial order A), whereas the first CS- trial was excluded for others (with trial order B; note that there were more participants in the 10 min group with trial order A than trial order B, and more participants with trial order B than trial order A in the 6 h/no-reminder group). We also identified a copy/pasting error in the final summary data file, which resulted in the shifting of some data points in the 6 h/no-reminder group. The latter error did not greatly affect the analyses, as the data were shifted within one and the same group only. Finally, and most critically, we found that rather than using the last 2 trials of re-extinction of the initial spontaneous recovery experiment (Day 3) to compare post-reinstatement responses to, Schiller et al. (2010) had used the last 2 trials of initial extinction (Day 2), unlike what was stated in their article. By doing the same, we were able to re-create the dataset Schiller and colleagues (2010) used for their reinstatement analysis (1-year FU).

After correcting all of the issues listed in the previous paragraph, we were able to replicate the main effect of group identically as reported ($F(1, 17) = 5.89$, $p = .027$, $\eta_p^2 = 0.26$). The critical interaction between group and time was not significant ($F(1, 17) = 2.78$, $p = .114$, $\eta_p^2 = 0.14$), but numerically identical to the F statistic reported by Schiller et al. (2010, p. 51) to be “one-tailed marginally significant” ($F(1, 17) = 2.78$, $p < .07$); see Discussion. Follow-up (one-tailed) t tests per group were likewise confirmed exactly; responding did not change significantly from the end of extinction to post-reinstatement in the 10 min group, whereas it increased in the 6 h/no-reminder group. Comparing the reinstatement index (again, using *extinction* rather than *re-extinction* data), we found no differences between groups, but we did obtain significant between-group differences in post-reinstatement SCR responding. For these last 2 t tests, Schiller and colleagues (2010) reported slightly different values and significance levels than what we obtained, apparently due to the fact that they used Welch’s t tests rather than Student t tests here, unlike in the rest of the article. In sum, when recreating the analyses actually conducted by Schiller and colleagues (rather than the analyses that they reported having conducted), we were able to reproduce (most of) their results. However, like for the spontaneous recovery results, the unprincipled exclusion of participants clearly shaped the pattern of results. Applying their reported exclusion criteria and/or stated statistical analysis plan rigorously eliminated any significant between-group differences (which were not statistically robust in the original analysis either).

3.6 Experiment 2 analyses

Differential responding increased from the first to the last half of acquisition (main effect of stimulus, $F(2, 34) = 10.85$, $p < .001$, $\eta_p^2 = 0.39$; stimulus x time, $F(2, 34) = 5.15$, $p = .011$, $\eta_p^2 = 0.23$), and declined from the first to the last half of extinction (main effect of time, $F(1, 17) = 30.04$, $p < .001$, $\eta_p^2 = 0.64$; stimulus x time, $F(1.35, 22.90) = 6.11$, $p = .014$, $\eta_p^2 = 0.26$). That pattern of results matches the one reported by Schiller et al. (2010), albeit with very different degrees of freedom, F values, and p values; their degrees of freedom also appear inconsistent with their experimental design and sample size. Follow-up t tests showed that during the last half of acquisition, participants responded more strongly to the CSa+ than the CS-, and the CSb+ than the CS-, while responding did not differ significantly between both CS+s (see Table 6 for an overview of all t test analyses in this section). By the last trial of extinction, there were no significant differences between any of the stimuli. Further, fear responding declined significantly from the end of acquisition to the last trial of extinction for the CSa+ and the CSb+, while it did not change significantly for the CS-.

With respect to the recovery of fear, unlike what Schiller et al. (2010) reported, we found that a 3 x 2 ANOVA comparing the first 4 to the subsequent 4 trials of re-extinction revealed no significant effects (main effect of stimulus, $F(2, 34) = 2.82$, $p = .074$, $\eta_p^2 = 0.14$; main effect of time, $F(1, 17) = 2.83$, $p = .11$, $\eta_p^2 = 0.14$; stimulus x time, $F(1.14, 19.33) = 1.01$, p

= .34, $\eta_p^2 = 0.06$). Follow-up *t* tests comparing the last trial of extinction to the beginning of re-extinction for each stimulus showed an increase in SCR responding only for the CSb+, while SCR did not recover significantly for the CSa+ or the CS- (see Figure 7). Note that Schiller et al. (2010) reported a *t*-value of 0.16 for the CS- comparison; tracing their steps we were able to decipher that they used the first CS- trial in their analysis, instead of excluding it as they reported.

In summary, while we were able to exactly replicate all of the *t* tests reported by Schiller et al. (2010, Experiment 2), we were not able to replicate any of the reported ANOVAs. For acquisition and extinction, we did observe similar effects, although with different degrees of freedom, *F* values, and *p* values. For the critical test of fear recovery, we were not able to confirm a significant stimulus by time interaction, unlike what was reported by the original authors.

4 Discussion

We set out to replicate the statistical analyses reported by Schiller et al. (2010) and to evaluate the influence of participant exclusions on the robustness of the reactivation-extinction effect as reported there. With respect to the spontaneous recovery results of their Experiment 1, in 4 separate sets of analyses the crucial pattern of a significant spontaneous recovery effect being absent in the reactivation-extinction group and present in the control groups was obtained only when considering the exact sample of participants selected by Schiller et al. (2010; see 3.1). Even then, no statistical support for a significant between-group difference in spontaneous recovery was obtained. In three analyses using more principled exclusion criteria, even less support for an advantage of reactivation-extinction over regular extinction in preventing spontaneous recovery was obtained. For the 1-year FU to their Experiment 1, we initially failed to reproduce the results reported by Schiller et al. (2010) and found no evidence for between-group differences in reinstatement. After identifying the discrepancies between how data processing and analysis were reported in Schiller et al. (2010) and how they were actually executed, we were able to mostly replicate their analysis, but still failed to find significant statistical support for a between-group difference in sensitivity to reinstatement. Finally, we were partially able to reproduce the analyses Schiller et al. (2010) reported for their Experiment 2; we could replicate the reported *t* tests but not their ANOVA results, and found in our ANOVA no support for a significant benefit of reactivation-extinction over regular extinction on the critical fear recovery test. Below we provide further observations regarding the analyses and results reported by Schiller et al. (2010) and discuss the implications of our findings and how they may illustrate the potential influence of experimenter decisions on the results of scientific research.

Schiller et al. (2010) consistently followed up non-significant ANOVA interactions using *t* tests examining time effects in each group separately, and then based their claims on those *t* tests. It is a matter of debate whether such follow-up analyses are appropriate in case of a non-significant omnibus ANOVA (Field, 2009; Nieuwenhuis, Forstmann, & Wagenmakers, 2011; Park, Cho, & Ki, 2009), but at any rate, they do not allow conclusions regarding group differences. Moreover, while some suggest that it is acceptable to perform multiple comparisons without an overarching significant ANOVA interaction, it is generally recommended to then correct for inflation of the type-I error rate through procedures such as Bonferroni adjustment or others (Hsu, 1996; Maxwell & Delaney, 2004).

Additionally, the follow-up statistical tests reported by Schiller et al. (2010) often do not match the ANOVAs they were intended to pursue. For example, in Experiment 1, acquisition and extinction were evaluated using rm-ANOVAs comparing the interaction of Group with the change in *differential SCR* responding over Time, across each phase (first half, last half). Yet, these analyses were followed up with *t* tests comparing *CS+ to CS-* responding (and not the differential SCR responding) on the last half (for acquisition) or last trial (for extinction) for each group separately. The more appropriate comparison given the preceding overall rm-ANOVAs would arguably have involved a comparison of the change in differential SCR from the first to the last half of each phase. Reduction of fear responding was evaluated by comparing differential responding on the last half of acquisition to that on the last trial of extinction, and this rm-ANOVA was appropriately followed-up using a *t* test of the same comparison within each group separately. Yet to evaluate spontaneous recovery, whereas (somewhat idiosyncratically) trials 1-4 and 5-8 of the re-extinction phase were used in the rm-ANOVA, this was followed up with (more typical) *t* tests comparing the last trial of extinction and the first trial of re-extinction per group. In general, it is unclear why the assessments in Experiment 1 (whether for acquisition, extinction, fear reduction, or spontaneous recovery) used variables for the ANOVAs that were different from the variables used in the follow-up tests, and different for distinct phases of the study. Moreover, for the reinstatement comparison of the 1-year FU, Schiller et al. (2010, p. 51) report the crucial group x time rm-ANOVA interaction to be “marginally significant ($F(1, 17) = 2.78$, $p < .07$, one-tailed)”, which fosters further questions regarding statistical choices and their reporting, given that *F* tests are always based on one tail only and that the *p* value of the interaction was actually .114.

Complicating matters further, Figure S1 in the supplementary information for Schiller et al. (2010) is inconsistent with Figure 1 in the main text. Whereas Figure 1 indicates that average differential responding was negative in all groups on the last trial of extinction in Experiment 1 (implying that CS- responding was higher than CS+ responding), Figure S1 suggests differently for the last trial of extinction in the no-reminder group (i.e., CS+ > CS-). In addition to this discrepancy, we observed further inconsistencies between the originally reported results (Schiller et al., 2010) and our re-analysis of the same dataset (see 3.1). Specifically, we found a difference between groups in the first half of the extinction phase of Experiment 1 that was not reported in Schiller et al. (2010), and different *t* values for follow-up analyses of the no-reminder group. To trace the origin of those discrepancies, we graphed the trial-by-trial SCR data of the 65 participants included in the original report, using the data available on the OSF (see Figure 4). By doing so, we were able to pinpoint the differences in the data that lead to the discrepancies between the original statistics and our re-analysis, as well as between the graphs within Schiller et al. (2010). When comparing our graphs in Figure 4 to Figure S1 of Schiller et al. (2010), it can be noticed that the data points of the first CS+ extinction trial for the 6 h group and of the last CS+ extinction trial for the no-reminder group do not match between the figures. This probably accounts for the conflicting results and raises questions as to what happened with the original dataset during processing. Perhaps slightly different trials were used for the supplementary graphs and the main analyses in Schiller et al. (2010), or perhaps the data made available on the OSF are slightly different from those used in the original report; these are issues that do not have a definitive solution at present. However, considering that our Figure 4 is in line with Figure 1 of Schiller et al. (2010), we tentatively conclude that for unknown reasons, they used slightly different data for some statistical comparisons and their Figure S1 than what they reported. The most notable inconsistency, however, and one that we did manage to resolve, is that Schiller et al. (2010) used distinctively different data (extinction rather than re-extinction data) and a different processing of the data (see 3.5) than what they reported for the 1-year FU test of reinstatement.

The use of responding-based participant exclusion criteria is not uncommon in human fear conditioning research, as it allows researchers to include only those participants who effectively acquire and extinguish a certain level of conditioned responding in experiments aimed to evaluate post-extinction manipulations. In keeping with this tradition, Schiller and colleagues (2018) listed a collection of inclusion criteria aimed at retaining only those participants who exhibited successful acquisition and extinction. However, the criteria reported for Experiment 1 are rather unconventional, in that they not only pertain to the end of the acquisition and extinction phases (as is common practice, if not unproblematic, see Lonsdorf et al., 2019), but also to various timepoints within and across phases (i.e., first half, last half, last trial, increase/decline in differentiation from first half/trial to last half/trial). Due to these idiosyncratic criteria, some participants not displaying any conditioned responding by the end of acquisition, or maintaining differential conditioned responding by the end of extinction, were nonetheless included (see Figure 3 for an example participant). Further, bearing in mind that these criteria were justified on the basis of the need for strong patterns of acquisition and extinction, it is remarkable that in our re-analyses of the original dataset, the strongest acquisition effect was observed in the analysis including all participants (see 3.4), where no exclusion criteria were applied. Lastly, it is important to point out that even though our four analyses generally yielded similar conclusions with respect to acquisition and extinction, the inclusion or exclusion of participants markedly influenced the pattern of SCR responding for the spontaneous recovery comparison (see Figure 5). In neither of the four sets of analyses did the omnibus ANOVAs produce a statistically significant interaction between Group and Time. Still, we followed them up with post-hoc *t* tests in line with the original analysis pipeline of Schiller et al. (2010). When the exclusion criteria were applied either meticulously or not at all, SCR responding did not significantly change from the end of extinction to the test phase in any of the groups. In contrast, when experimenter decisions were made, and thus exclusions were qualitative (see 3.1 and Figure 5A), significant spontaneous recovery remained absent in the 10 min group but was now present in the other two groups.

Qualitative evaluations also appear to have influenced the results of the long-term reinstatement FU of Experiment 1. Schiller et al.'s (2010) basis for their decision to exclude 5 participants from their analysis, of which only 1 fit into the exclusion criteria as reported (i.e., as being a non-responder), remains unclear. With those exclusions, and recreating the statistical comparisons actually conducted by Schiller et al. (2010), reinstatement cannot be demonstrated in the reactivation-extinction group whereas it is present in the control group, although the effect is not significantly different between the groups.

For Experiment 2, it is more difficult to reach conclusions regarding the possible influence of qualitative participant exclusions on the results. Similar to Experiment 1, “inclusion of participants was based on qualitative assessment of trial-by-trial SCR” (Schiller et al., 2018, p. 7), and only the data of the 18 included participants were made available on the OSF. We did however find data for an additional 6 excluded participants in the files that we received directly from the authors. Whereas according to Schiller et al. (2018, p. 7), participants were excluded “…for failing to acquire or extinguish conditioned responses to both CS+ stimuli”, it is unclear whether that was indeed the basis for the exclusion of those 6 participants, in view of the data of participants who were included. For example, comparing the last half of acquisition data for two of the excluded participants (1: CSa+ = 0.25, CSb+ = 0.22, CS- = 0.18; 2: CSa+ = 0.42, CSb+ = 0.44, CS- = 0) to the last half of acquisition data for two of the included participants (1: CSa+ = 0.48, CSb+ = 0.09, CS- = 0.07; 2: CSa+ = 0.02, CSb+ = 0.21, CS- = 0.005), one might think that the included participants should have been excluded and vice versa. At any rate, also here, no evidence was obtained for a significant difference in fear recovery between a reactivated-and-extinguished cue and a merely extinguished cue.

In this report, we have tried to evaluate the veracity and robustness of the seminal findings of Schiller et al. (2010). Our observations highlight the potential impact of experimenter decisions on experimental results and their subsequent interpretation. Schiller and colleagues (2010) did not employ their originally reported exclusion criteria, nor did they rigorously apply the ones later reported in their addendum (Schiller et al., 2018). They did not accurately report the total number of participants tested in their initial paper (Schiller et al., 2010) and their exclusions lacked justification (Schiller et al., 2018). Lastly, they presented incorrect graphs, misreported the data points they used for some of their crucial analyses, and conducted incoherent combinations of ANOVAs and *t* tests (Schiller et al., 2010). It is clear from the re-analyses presented here that their key observations do not stand when these inconsistencies are removed. In light of those observations, we conclude that the findings reported by Schiller et al. (2010) are unreliable and flawed. In further support of this conclusion, in a high-powered, direct replication of Schiller et al. (2010, Experiment 1), we failed to observe any benefit of reactivation-extinction over regular extinction training when using the exact same protocol and procedures (Chalkia et al., submitted). We therefore believe that, when evaluating the current evidence for and against the existence and robustness of the reactivation-extinction effect, the findings of Schiller et al. (2010) should not be taken into consideration.

More broadly, this verification report should serve as a demonstration of how experimenter influence can bias scientific findings. Flexibility in experiment planning, data collection, analysis and reporting, often referred to as *experimenter degrees of freedom*, carries a heavy risk of yielding false positive results that bias the scientific record (Wicherts et al., 2016). In order to prevent the infiltration of the scientific literature with reports that are not up to par, Simmons and colleagues (2011) provided some guidelines for authors and reviewers. For authors, these included simple steps such as describing decisions for the amount of observations that will be collected and full reporting of all exclusions, variables, conditions, and analyses. In the wake of their suggestions, new initiatives emerged, such as tools for pre-registration and the establishment of registered reports, which aspire to shift the focus of researchers from a need to publish to a need to do solid science (Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek & Lakens, 2014). Methods, materials and data sharing, as well as pre-registration of designs, hypotheses, and statistical analysis plans are only some of the actions that have been recommended (Nosek et al., 2015; Nosek, Spies, & Motyl, 2012). As we look into the future, the promotion of open, transparent and reproducible science should be on every researcher's mind as a positive and promising way forward.

Acknowledgements

The authors would like to thank Angelos Miltiadis Krypotos for his help with the Bayesian analyses and their interpretation.

Funding

This work was supported by ERC Consolidator Grant 648176 awarded to Tom Beckers. The funding source was not involved in the preparation of this manuscript.

References

Auber A, Tedesco V, Jones CE, Monfils MH, Chiamulera C. Post-retrieval extinction as reconsolidation interference: Methodological issues or boundary conditions? Psychopharmacology. 2013; 226(4):631–647. DOI: 10.1007/s00213-013-3004-1 [PubMed: 23404065]

Bouton ME. Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. Biological Psychiatry. 2002; 52(10):976–986. DOI: 10.1016/S0006-3223(02)01546-9 [PubMed: 12437938]

Bouton ME. Context and behavioral processes in extinction. Learning & Memory. 2004; 11(5):485–494. DOI: 10.1101/lm.78804 [PubMed: 15466298]

Braithwaite JJ, Watson DG, Jones R, Rowe M. A Guide for Analysing Electrodermal Activity (EDA) and Skin Conductance Responses (SCRs) for Psychological Experiments …. 2015:1–42.

Chalkia, A, Schroyens, N, Leng, L, Vanhasbroeck, N, Zenses, A-K, Van Oudenhove, L, Beckers, T. No persistent attenuation of fear memories in humans: A registered replication of the reactivation-extinction effect. Manuscript submitted for publication; 2019.

Field, A. Discovering statistics using IBM SPSS statistics. 3rd ed. London, United Kingdom: Sage; 2009.

Hsu, JC. Multiple comparisons: Theory and methods. London, United Kingdom: Chapman & Hall; 1996.

JASP Team. JASP (Version 0.9.2). 2019

Jeffreys, H. Theory of probability. 3rd ed. Oxford, United Kingdom: Clarendon Press; 1961.

Kredlow MA, Unger LD, Otto MW. Harnessing reconsolidation to weaken fear and appetitive memories: a meta-analysis of post-retrieval extinction effects. Psychological Bulletin. 2016; 142(3):314–336. DOI: 10.1037/bul0000034 [PubMed: 26689086]

Lee JLC. Reconsolidation: Maintaining memory relevance. Trends Neurosci. 2009; 32(8):413–420. DOI: 10.1016/j.tins.2009.05.002.Reconsolidation [PubMed: 19640595]

Lonsdorf TB, Klingelhöfer-Jens M, Andreatta M, Beckers T, Chalkia A, Gerlicher A, Jentsch VL, Meir Drexler S, Mertens G, Richter J, Sjouwerman R, et al. Navigating the garden of forking paths for data exclusions in fear conditioning research. eLife. 2019; 8:e52465.doi: 10.7554/eLife.52465 [PubMed: 31841112]

Maxwell, SE, Delaney, HD. Designing experiments and analyzing data: A model comparison perspective. 2nd ed. New Jersey, USA: Lawrence Erlbaum Associates Publishers; 2004.

Monfils MH, Cowansage KK, Klann E, LeDoux JE. Extinction-reconsolidation boundaries: Key to persistent attenuation of fear memories. Science. 2009; 324(5929):951–955. DOI: 10.1126/science.1167975 [PubMed: 19342552]

Monfils MH, Holmes EA. Memory boundaries: opening a window inspired by reconsolidation to treat anxiety, trauma-related, and addiction disorders. The Lancet Psychiatry. 2018; 5(12):1032–1042. DOI: 10.1016/S2215-0366(18)30270-0 [PubMed: 30385214]

Nieuwenhuis S, Forstmann B, Wagenmakers E. Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience. 2011; 14:1105–1107. DOI: 10.1038/nn.2886 [PubMed: 21878926]

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Yarkoni T. Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science. 2015; 348(6242):1422–1425. DOI: 10.1126/science.aab2374.Promoting [PubMed: 26113702]

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. PNAS. 2018; 115(11):2600–2606. DOI: 10.1073/pnas.1708274114 [PubMed: 29531091]

Nosek BA, Lakens D. Registered Reports: A method to increase the credibility of published results. Social Psychology. 2014; 45(3):137–141. DOI: 10.1027/1864-9335/a000192

Nosek BA, Spies JR, Motyl M. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science. 2012; 7(6):615–631. DOI: 10.1177/1745691612459058 [PubMed: 26168121]

Park E, Cho M, Ki C-S. Correct use of repeated measures analysis of variance. Korean Journal of Laboratory Medicine. 2009; 29:1–9. DOI: 10.3343/kjlm.2009.29.1.1 [PubMed: 19262072]

Przybyslawski J, Sara SJ. Reconsolidation of memory after its reactivation. Behavioural Brain Research. 1997; 84:241–246. DOI: 10.1016/S0166-4328(96)00153-2 [PubMed: 9079788]

Sara SJ. Retrieval and reconsolidation: Toward a neurobiology of remembering. Learning & Memory. 2000; 7(2):73–84. DOI: 10.1101/lm.7.2.73 [PubMed: 10753974]

Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA. Preventing the return of fear in humans using reconsolidation update mechanisms. Nature. 2010; 463(7277):49–53. DOI: 10.1038/nature08637 [PubMed: 20010606]

Schiller D, Monfils MH, Raio CM, Johnson DC, LeDoux JE, Phelps EA. Addendum: Preventing the return of fear in humans using reconsolidation update mechanisms. Nature. 2018; 562(7727):E21.doi: 10.1038/s41586-018-0405-7 [PubMed: 30050064]

Schiller D, Raio CM, Phelps EA. Extinction training during the reconsolidation window prevents recovery of fear. Journal of Visualized Experiments: JoVE. 2012; 66:e3893.doi: 10.3791/3893

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22(11):1359–1366. DOI: 10.1177/0956797611417632 [PubMed: 22006061]

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. Perspectives on Psychological Science. 2011; 6(3):291–298. DOI: 10.1177/1745691611406923 [PubMed: 26168519]

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Frontiers in Psychology. 2016; 7:1832.doi: 10.3389/fpsyg.2016.01832 [PubMed: 27933012]

Zuccolo PF, Hunziker MHL. A review of boundary conditions and variables involved in the prevention of return of fear after post-retrieval extinction. Behavioural Processes. 2019; 162:39–54. DOI: 10.1016/j.beproc.2019.01.011 [PubMed: 30708059]

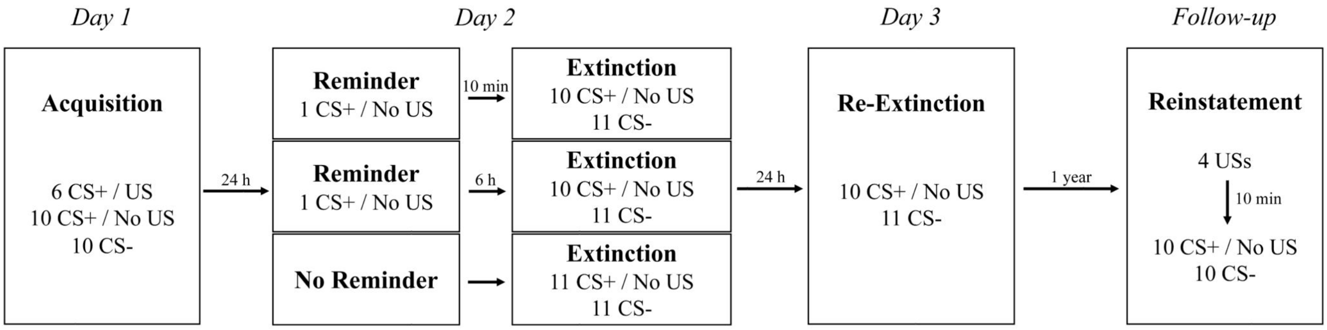

Figure 1.

Schematic overview of Schiller et al. (2010, Experiment 1).

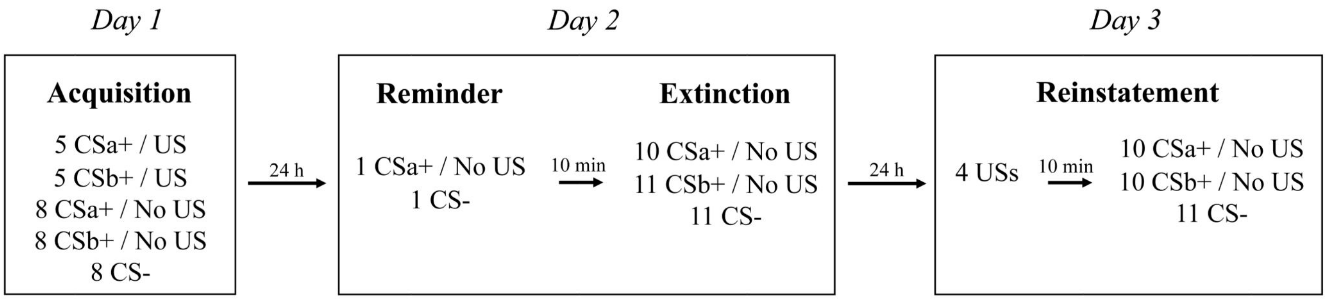

Figure 2.

Schematic overview of Schiller et al. (2010, Experiment 2).

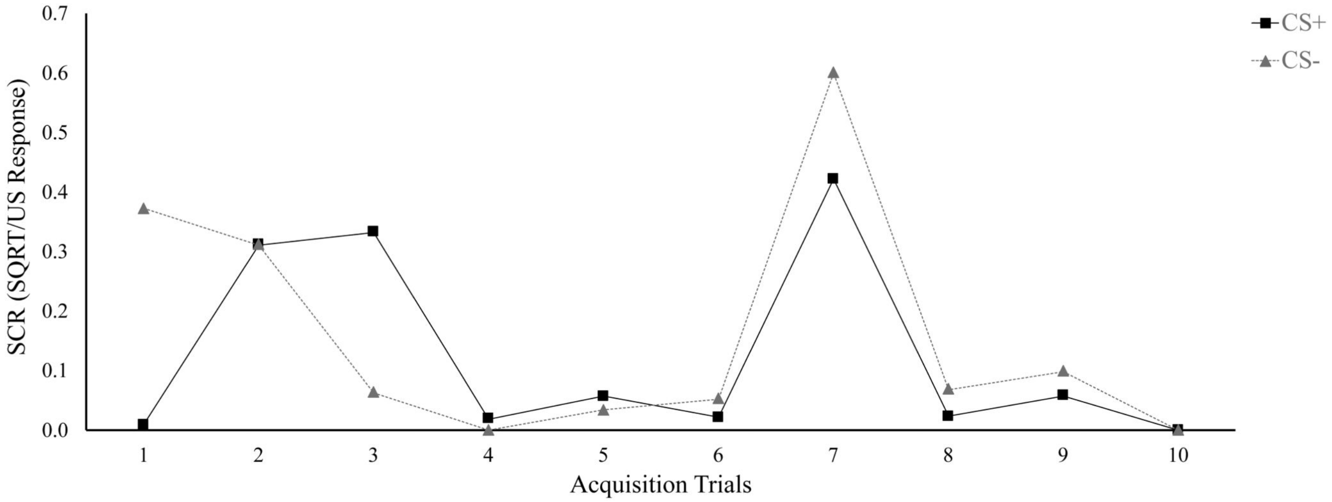

Figure 3.
Example acquisition data from Schiller et al. (2010, Experiment 1) of a participant that displayed no differential conditioned responding yet met the provided criteria for acquisition.

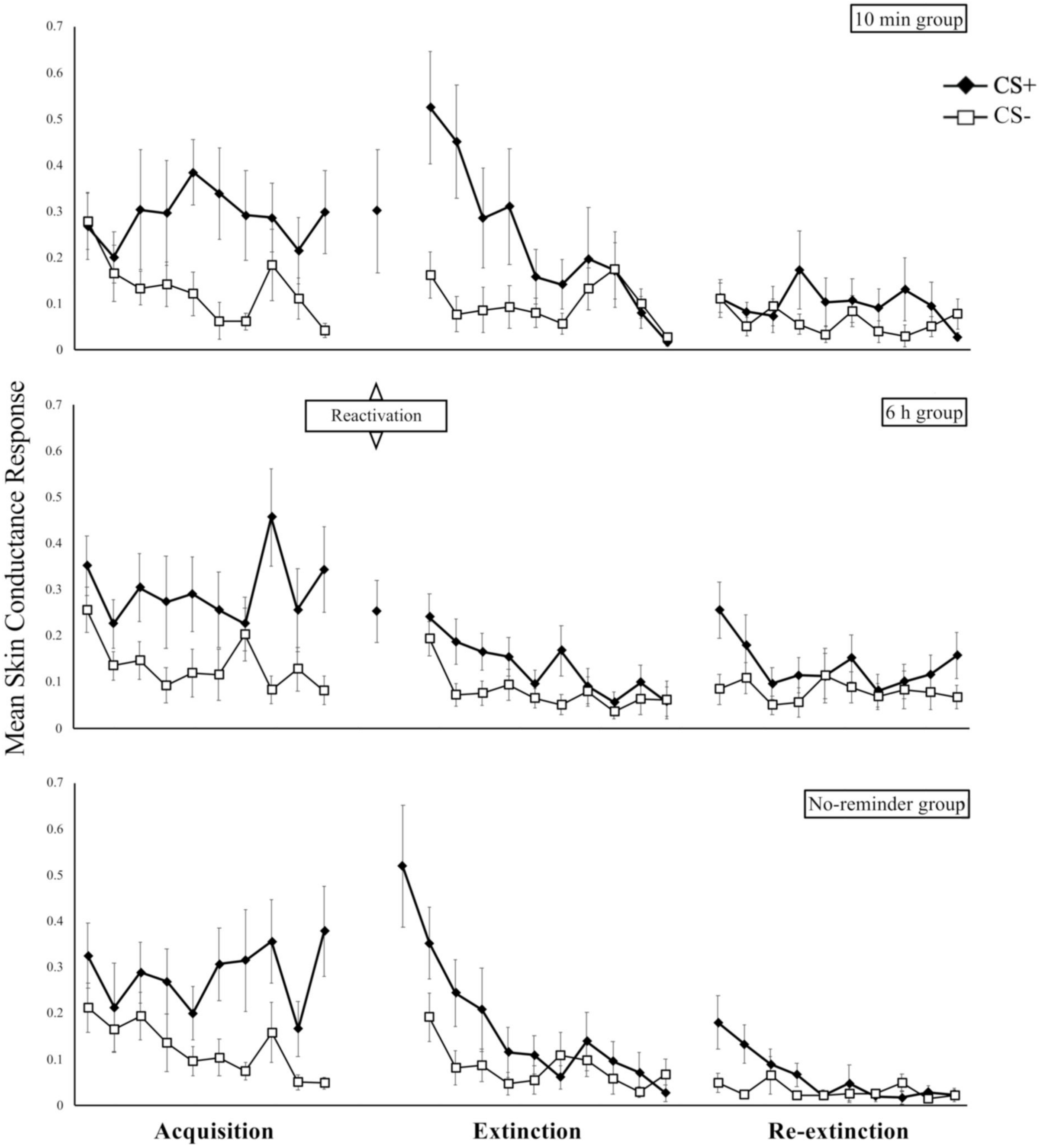

Figure 4.

Trial-by-trial SCR data of the 65 participants included by Schiller et al. (2010). Error bars represent standard error of the mean.

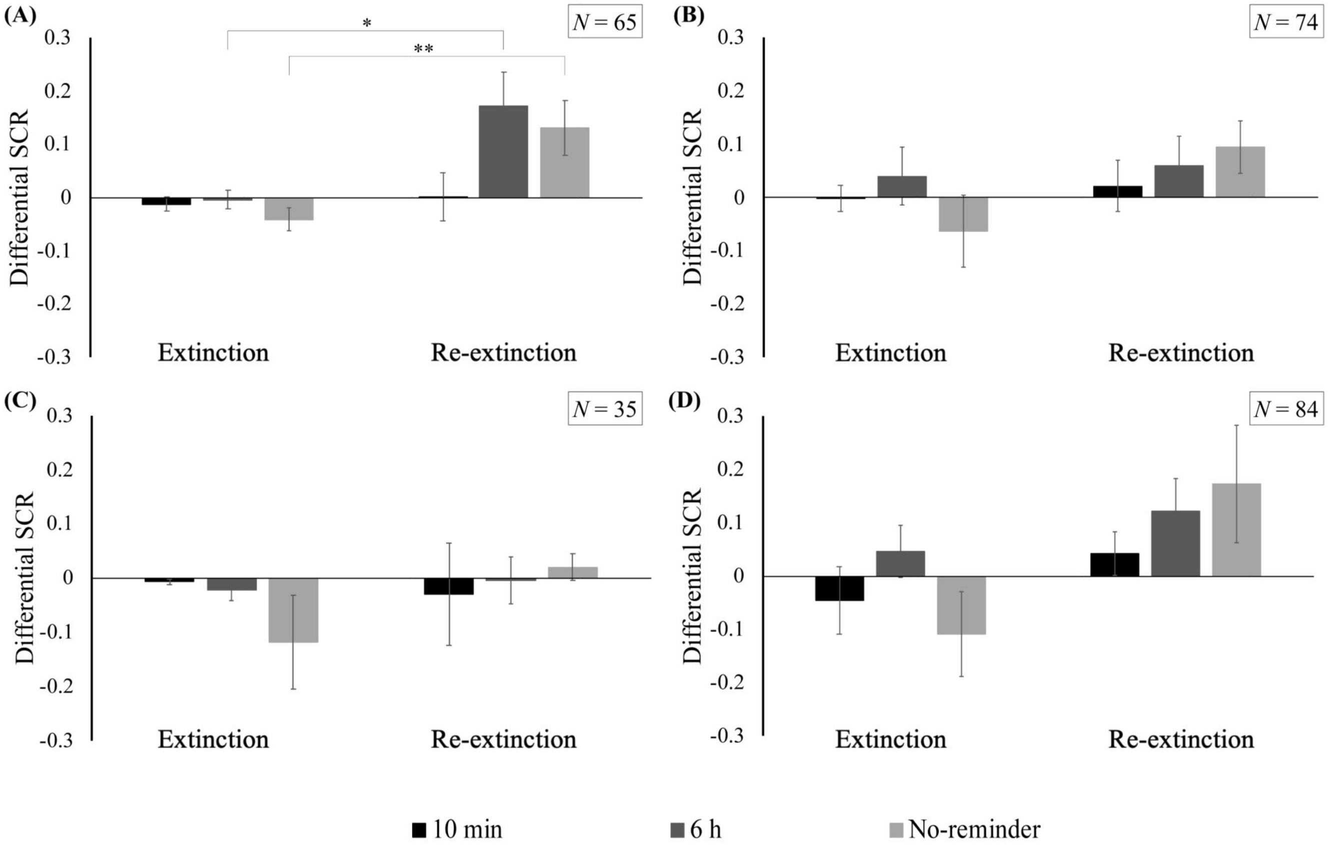

Figure 5.

Differential SCR responding on the last trial of extinction and the first trial of re-extinction for **(A)** Analysis 3.1 [*N* = 65; final sample included by Schiller et al. (2010)], **(B)** Analysis 3.2 [*N* = 74; inclusions based on criteria reported in Schiller et al. (2018)], **(C)** Analysis 3.3 [*N* = 35; inclusions based on initial criteria reported in Schiller et al. (2010)], and **(D)** Analysis 3.4 [*N* = 84; full sample]. Error bars represent standard error of the mean. *$p < .05$, **$p < .01$

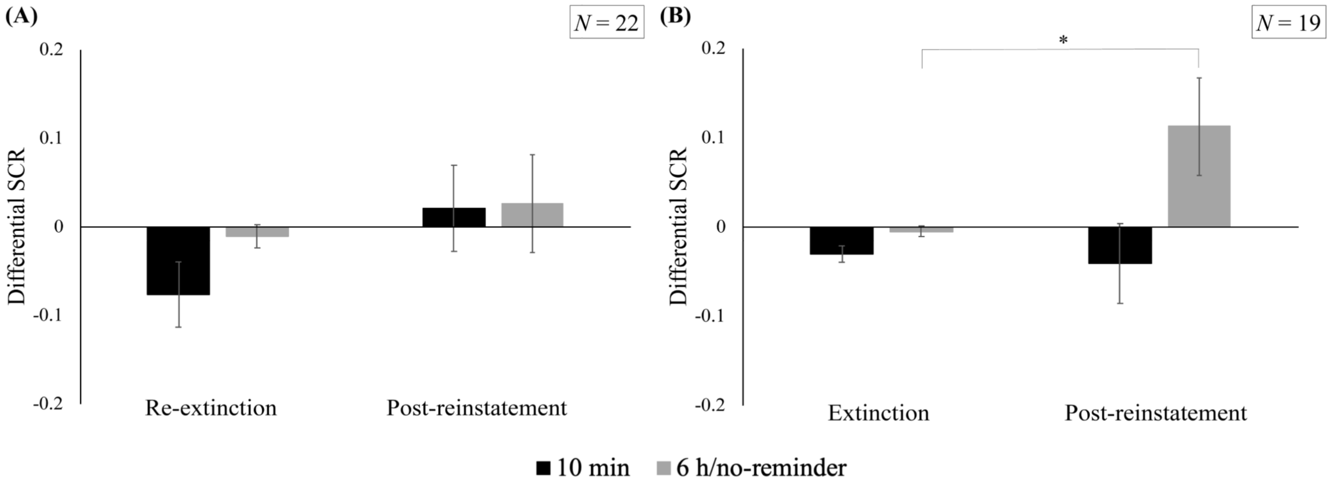

Figure 6.

Differential SCR responding during (re-)extinction (last 2 trials) and post-reinstatement (first 4 trials) for **(A)** Analysis with *N* = 22; applying criteria reported in *Nature*, and **(B)** Analysis with *N* = 19; applying choices made by the original authors. Error bars represent standard error of the mean. *$p < .05$

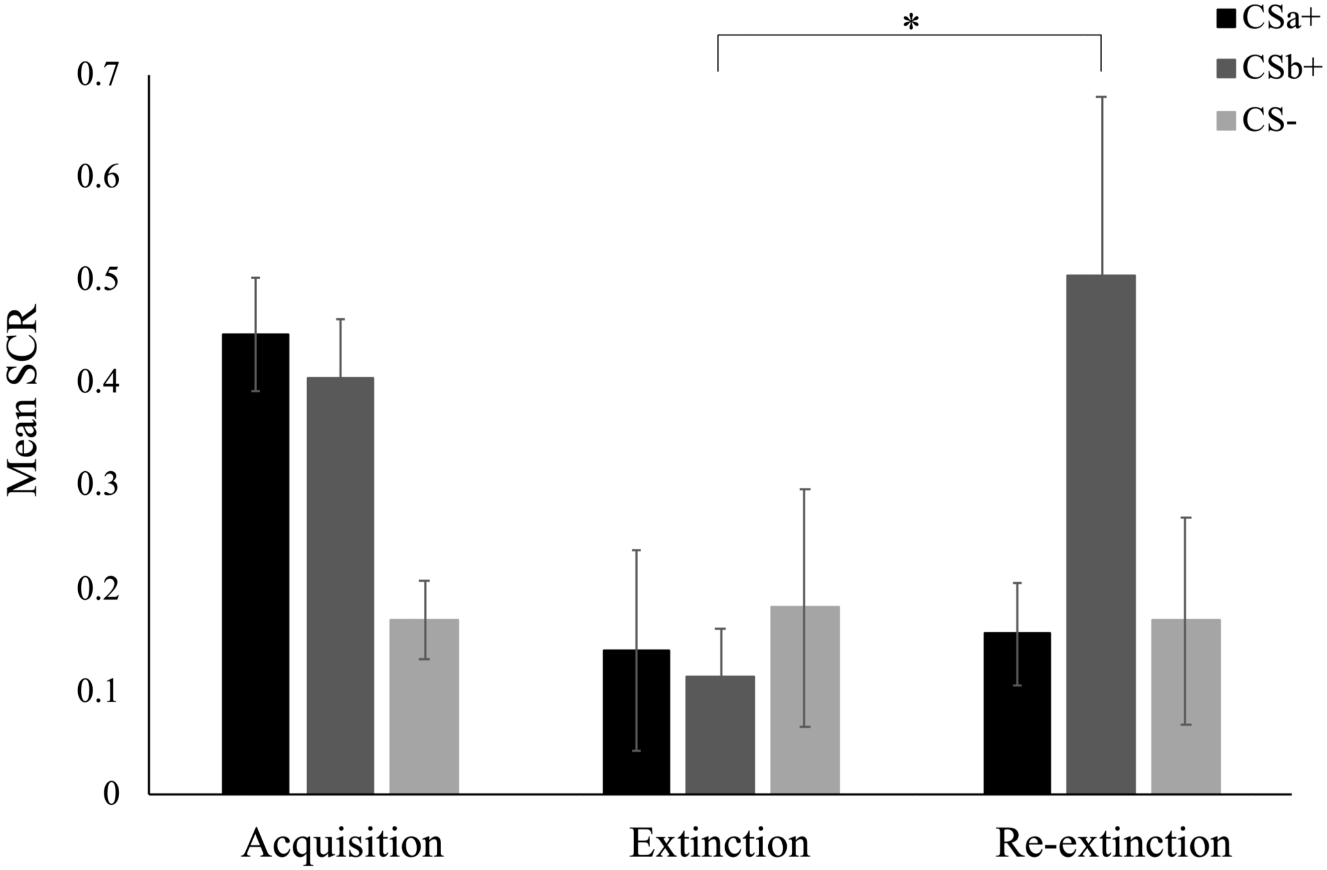

Figure 7.

Mean standardized SCR responding during acquisition (last half), extinction (last trial), and re-extinction (first trial) for each stimulus. Error bars represent standard error of the mean. $*p < .05$

Table 1

Mean values (SD) for standardized SCR, by group, and corresponding paired-sample t tests, for the analyses in section 3.1 (N = 65)

		Groups		
		10 min	**6 h**	**No-reminder**
Acquisition	Last half (CS+)	0.29 (0.31)	0.31 (0.29)	0.31 (0.32)
	Last half (CS-)	0.09 (0.13)	0.12 (0.17)	0.09 (0.11)
	t	2.68^{*}	3.71^{***}	3.72^{***}
Extinction	Last trial (CS+)	0.02 (0.03)	0.06 (0.15)	0.03 (0.09)
	Last trial (CS-)	0.03 (0.04)	0.06 (0.20)	0.07 (0.16)
	t	0.94	0.23	1.87
Fear reduction	Last half acquisition	0.19 (0.32)	0.19 (0.24)	0.22 (0.27)
	Last trial extinction	-0.01 (0.06)	-0.004 (0.08)	-0.04 (0.10)
	t	2.70^{*}	4.06^{***}	4.35^{***}
Spontaneous recovery	Last trial extinction	-0.01 (0.06)	-0.004 (0.08)	-0.04 (0.10)
	First trial re-extinction	0.002 (0.21)	0.17 (0.31)	0.13 (0.24)
	t	0.28	2.66^{*}	3.25^{**}

When CS+/CS- are not indicated, differential scores are reported (CS+ minus CS-). $^{*}p < .05$, $^{**}p < .01$, $^{***}p \leq .001$

Table 2

Mean values (SD) for standardized SCR, by group, and corresponding paired-sample t tests, for the analyses in section 3.2 (N = 74)

		Groups		
		10 min	**6 h**	**No-reminder**
Acquisition	Last half (CS+)	0.32 (0.31)	0.26 (0.20)	0.33 (0.30)
	Last half (CS-)	0.11 (0.14)	0.13 (0.17)	0.12 (0.14)
	t	2.95^{**}	5.89^{***}	4.56^{***}
Extinction	Last trial (CS+)	0.09 (0.23)	0.12 (0.30)	0.10 (0.23)
	Last trial (CS-)	0.10 (0.22)	0.08 (0.20)	0.16 (0.39)
	t	0.06	0.73	0.92
Fear reduction	Last half acquisition	0.20 (0.32)	0.14 (0.11)	0.21 (0.25)
	Last trial extinction	-0.001 (0.11)	0.04 (0.27)	-0.06 (0.37)
	t	2.65^{*}	1.55	3.46^{**}
Spontaneous recovery	Last trial extinction	-0.001 (0.11)	0.04 (0.27)	-0.06 (0.37)
	First trial re-extinction	0.02 (0.22)	0.06 (0.27)	0.10 (0.27)
	t	0.40	0.25	1.92

When CS+/CS- are not indicated, differential scores are reported (CS+ minus CS-). *p < .05, **p < .01, ***p < .001

Table 3

Mean values (SD) standardized SCR, by group, and corresponding paired-sample t tests, for the analyses in section 3.3 (N = 35)

		Groups		
		10 min	**6 h**	**No-reminder**
Acquisition	Last half CS+	0.27 (0.34)	0.36 (0.22)	0.31 (0.33)
	Last half CS-	0.03 (0.02)	0.15 (0.25)	0.10 (0.11)
	t	2.22	7.81 ***	3.40 **
Extinction	Last trial (CS+)	0.01 (0.01)	0.02 (0.04)	0.02 (0.03)
	Last trial (CS-)	0.01 (0.02)	0.04 (0.10)	0.13 (0.36)
	t	1.19	1.12	1.36
Fear reduction	Last half acquisition	0.25 (0.33)	0.21 (0.08)	0.22 (0.25)
	Last trial extinction	-0.01 (0.02)	-0.02 (0.06)	-0.12 (0.35)
	t	2.28	8.42 ***	2.89 **
Spontaneous	Last trial extinction	-0.01 (0.02)	-0.02 (0.06)	-0.12 (0.35)
recovery	First trial re-extinction	-0.03 (0.28)	-0.004 (0.14)	0.02 (0.10)
	t	0.24	0.41	1.36

When CS+/CS- are not indicated, differential scores are reported (CS+ minus CS-). **p ≤.01, ***p < .001

Table 4

Mean values (SD) for standardized SCR, by group, and corresponding paired-sample t tests, for the analyses in section 3.4 (N = 84)

		Groups		
		10 min	**6 h**	**No-reminder**
Acquisition	Last half CS+	0.34 (0.30)	0.30 (0.27)	0.35 (0.32)
	Last half CS-	0.13 (0.16)	0.13 (0.16)	0.15 (0.23)
	t	3.43 **	4.02 ***	4.49 ***
Extinction	Last trial (CS+)	0.11 (0.24)	0.12 (0.29)	0.13 (0.27)
	Last trial (CS-)	0.15 (0.32)	0.07 (0.19)	0.24 (0.57)
	t	0.72	0.95	1.36
Fear reduction	Last half acquisition	0.21 (0.31)	0.17 (0.22)	0.20 (0.25)
	Last trial extinction	-0.05 (0.32)	0.05 (0.25)	-0.11 (0.45)
	t	3.30 **	1.98	3.65 ***
Spontaneous recovery	Last trial extinction	-0.05 (0.32)	0.05 (0.25)	-0.11 (0.45)
	First trial re-extinction	0.04 (0.21)	0.12 (0.31)	0.17 (0.62)
	t	1.21	0.97	1.68

When CS+/CS- are not indicated, differential scores are reported (CS+ minus CS-). **p < .01,***p ≤.001

Table 5

Mean values (SD) for standardized SCR, by group, and corresponding t tests, for the analyses in section 3.5

	Groups (*N* = 22)			**Groups (*N* = 19)**		
	10 min	**6 h/no-reminder**	***t***	**10 min**	**6 h/no-reminder**	***t***
Re-extinction	-0.08 (0.10)	-0.01 (0.05)		-0.03 (0.03)	-0.01 (0.02)	
Post-reinstatement	0.02 (0.14)	0.03 (0.21)	0.06	-0.04 (0.13)	0.11 (0.18)	2.06 *
t	1.69	0.60		0.22	2.12 *	
Reinstatement index	0.10 (0.16)	0.04 (0.22)	0.67	-0.01 (0.14)	0.12 (0.18)	1.67

Mean values include the last 2 trials for re-extinction and the first 4 trials for post-reinstatement; the difference between post-reinstatement and re-extinction yields the reinstatement index. Paired t tests compared re-extinction to post-reinstatement per group. Independent t tests compared post-reinstatement and the reinstatement index between groups. Note that for the right panel “Groups (N = 19)” extinction data (last 2 trials) are reported instead of re-extinction data, following the apparent processing of the original authors (see text for details). *p < .05

Table 6

Mean values (SD) for standardized SCR, by stimulus, and corresponding paired-sample t tests, for the analyses in section 3.6

		Stimuli			
		CSa+	**CSb+**	**CS-**	***t***
Acquisition	*M* (*SD*)	0.45 (0.24)	0.40 (0.25)	0.17 (0.16)	
	CSa+ vs. CS-				6.01^{***}
	CSb+ vs. CS-				6.68^{***}
	CSa+ vs. CSb+				0.76
Extinction	*M* (*SD*)	0.14 (0.41)	0.11 (0.20)	0.18 (0.49)	
	CSa+ vs. CS-				0.26
	CSb+ vs. CS-				0.56
	CSa+ vs. CSb+				0.23
Fear reduction	CSa+				2.62^{*}
	CSb+				4.08^{***}
	CS-				0.09
Re-extinction	*M* (*SD*)	0.16 (0.21)	0.51 (0.74)	0.17 (0.17)	
	CSa+				0.22
	CSb+				2.16^{*}
	CS-				0.08

Mean values include the last half for acquisition, the last trial for extinction and the first trial for re-extinction. T tests for acquisition and extinction compared each stimulus to the other 2 stimuli. For fear reduction and re-extinction, acquisition was compared to extinction, and extinction to re-extinction, respectively, for each stimulus separately. *p < .05, ***p < .001