Running head: COUNTERCONDITIONING AND EXTINCTION

Retroactive Interference:

Counterconditioning and Extinction with and without Biologically Significant Outcomes

Jérémie Jozefowiez,

Université de Lille

Alaina S. Berruti, Yaroslav Moshchenko, Tori Peña, Cody W. Polack, & Ralph R. Miller

SUNY - Binghamton

Correspond:
Dr. Jérémie Jozefowiez
Sciences Cognitives et Affectives (SCALab) – UMR 9193
Université de Lille
Campus de Lille 3
Domaine Universitaire du Pont de Bois
B.P. 60149
59653 Villeneuve d'Ascq Cedex
France
Email: jeremie.jozefowiez@univ-lille.fr
Phone: +33-320-041-6866
Fax: +33-320-041-6036

Submitted: December 2019; revised June 2020

Abstract

Following cue-outcome (X-O) pairings, two procedures that reduce conditioned responses to X are extinction, in which X is presented by itself, and counterconditioning, in which X is paired with a different outcome typically of valence opposite that of training. While studies with animals have generally found counterconditioning more efficient than extinction in reducing responding, data from humans are less clear. They suggest counterconditioning is more efficient than extinction at interfering with emotional processing, but there is little difference between the two procedures regarding their impact on the verbal assessment of the probability of the outcome given the cue. However, issues of statistical power leave conclusions ambiguous. We compared counterconditioning and extinction in highly powered experiments that exploited a novel procedure. A rapid streamed-trial procedure was used in which participants were asked to rate how likely a target outcome was to accompany a target cue after being exposed to acquisition trials followed by extinction, counterconditioning, or neither. In Experiments 1 and 2, evaluative conditioning was assessed by asking participants to rate the pleasantness of the cues after treatment. These studies found counterconditioning more efficient than extinction at reducing evaluative conditioning but less efficient at decreasing the assessment of the conditional probability of the outcome given the cue. The latter effect was replicated with neutral outcomes in Experiments 3 and 4, but the effect was inverted in Experiment 4 in conditions designed to preclude reinstatement of initial training by the question probing the conditional probability of the outcome given the cue. Effect sizes were small (Cohen's *d* of 0.2 for effect on evaluative conditioning, Cohen's *d* of 0.3 for effect on the outcome expectancy). If representative, this poses a serious constraint in terms of statistical power for further investigations of differential efficiency of extinction and counterconditioning in humans.

Keywords: associative learning, evaluative conditioning, extinction, counterconditioning, rapid streamed-trial procedure.

In Pavlovian conditioning, an initially neutral [to be] conditioned stimulus (CS, a.k.a. a cue) is paired with a biologically relevant unconditioned stimulus (US) with the consequence that the CS comes to elicit a US-appropriate conditioned response (CR). This is often thought to reflect the creation of an association between internal representations of the CS and the US. Though typically adaptive, the associative process underlying the development of the CR can sometimes result in maladaptive behavior. For instance, a leading theory of anxiety disorders and phobias regard these disorders as CRs triggered by a CS previously paired with an aversive US (e.g., VanElzakker, Dahlgren, Davis, Dubois, & Shin, 2014; Vervliet, Craske, & Hermans, 2013). As a consequence, considerable effort has been invested in trying to uncover the most efficient way of reducing maladaptive conditioned responding.

The two main procedures used to achieve this goal are extinction and counterconditioning (CC; Bouton, 2017). In extinction, the CS is simply presented by itself, unaccompanied by the original US (US1), whereas in CC, the CS is paired with another US (US2), the emotional valence of which is often the opposite of the one used to condition the CS. For instance, in therapy, a patient would be asked to think of a relaxing memory (US2) when presented with an anxiety-inducing stimulus (the CS). In principle, one might expect CC to be more effective than extinction at reducing CRs because by administering CS-US2 trials, not only is the CS presented without US1 (CS-noUS1) as in extinction treatment, but also with a distinctly different outcome (US2) which could further interfere with expression of the CS-US1 memory beyond that of only CS-noUS1 treatment. Alternatively, it is also possible that during CC, CS-US2 conditioning could impede learning about the absence of US1 because learning processes might be preoccupied with learning about US2. Rather than further speculate about the relative efficacies of extinction and CC, we turn to data.

While neither extinction nor CC seems to erase the CS-US1 association as evidenced by their susceptibility to spontaneous recovery, reinstatement, and renewal (Bouton, 2017), they are effective in decreasing the potential of the CS to elicit a CR. Research on non-human animals (Escobar, Arcediano, & Miller, 2001; Holmes, Leung, & Westbrook, 2016; Tunstall, Verendeev, & Kearns, 2012) has found that CC is more efficient than extinction at reducing conditioned responding.

Dunsmoor et al. (2015) found no difference between extinction and CC in a fear conditioning procedure with rats. Importantly, they used a neutral outcome during CC instead of an outcome whose emotional valence is the reverse of the one used during acquisition. The conclusions from research on human participants are not so straightforward. Some studies report that CC is more efficient than extinction at reducing the emotional responses triggered by a CS (Engelhard, Leer, Lange, & Olatunji, 2014; Kerkhof, Vansteenwegen, Baeyens, & Hermans, 2011; Reynolds, Fields, & Askew, 2018). Dunsmoor et al. (2019), presenting a neutral outcome immediately after the CS presentations during CC (which they called "enhanced extinction" treatment rather than CC due to the nontarget outcome in Phase 2 not being biologically relevant) also reported that this sort of counterconditioning was more efficient than extinction, while other studies concluded that there is no difference between the two treatments in this regard (De Jong, Vorage, & Van den Hout, 2000; Kang, Vervliet, Engelhard, van Dis, & Hagenaars, 2018; Meulders, Karsdop, Claes, & Vlaeyen, 2015; Raes & De Raedt, 2012). Dunsmoor et al. (2015) and Lucas, Luck, and Lipp (2018), both using a neutral outcome during CC, also reported a failure to find a difference between CC and extinction. In contrast, other studies have consistently failed to detect any difference between CC and extinction with respect to the prediction of the US1 in presence of the CS (Engelhard et al., 2014; Meulders et al., 2015; Raes & De Raedt, 2012), with the exception of Kang et al. (2018) which concluded that CC might be somewhat faster at reducing US1 expectancy than is extinction.

It is unclear how to interpret the divergent results reported in the human literature. The effects reported in the non-human research studies are quite large in that statistical tests have proven statistically significant despite the small number of animals observed. This suggests that comparable effects should have been readily detected in the human data if the effects were of similar magnitude across species. However, this is not the case, which leaves us without guidance regarding how large a potential difference between CC and extinction might be expected and hence how many participants are needed to achieve reasonable statistical power.

The median number of participants per group for the human studies mentioned above was 24 for those that only recorded emotional measures, and 33 for those that have looked at US expectancy

in addition to emotional measures. An effect size corresponding to a Cohen's *d* of 0.5 is supposed to be quite typical in psychology (a medium-size effect according to Cohen's 1988 guidelines). In the absence of better information about effect size in humans, this should be the prior. In an independent design, statistical power for a t-test barely reaches 51% for 33 participants per group whereas it drops to 39% for 24 participants per group. If the difference between CC and extinction is smaller than the typical effect size, differences will become more difficult to detect. For instance, for a Cohen's *d* of 0.4 (a Cohen's *d* of 0.3 is a small size effect according to Cohen's 1988 guidelines), statistical power for a t-test drops to 36% with 33 participants per group and 27% for 24 participants per group. Hence, it is likely that most of the human studies which have contrasted CC and extinction were underpowered, which would explain the mixed pattern of results and highlight the need for a high-powered study designed to reveal differences, if any, between CC and extinction. This is what we aimed to achieve in the present series of experiments.

Participants were exposed to rapid streams of trials, based on the streamed trial procedure of Lorraine Allan and her colleagues (Crump, Hannah, Allan, & Hord, 2007; Hannah, Crump, Allan, & Siegel, 2009; Laux, Goedert, & Markman, 2010; Maia, Lefèvre, & Jozefowiez, 2018; Siegel, Crump & Allan, 2009). In Phase 1, two stimuli, X (the cue) and O1 (the outcome), were simultaneously paired in the stream of trials. In Phase 2 X then appeared by itself (Extinction condition), simultaneously paired with another outcome O2 (CC condition), or not at all (Control condition). Finally, participants were asked the probability that the outcome would appear if the cue was presented next (expectancy rating). The valence of the outcome was also assessed at the end of the experiment. Note that our use 'cue' and 'outcome' terminology here is based on the wording of the expectancy rating question (the participant is asked about the likelihood of the outcome conditional on the cue) despite the simultaneous onset of the so-called cue and so-called outcome during training pairings. Specifically, participants were asked about the actual images of the stimuli witnessed during training, and not the terms 'cue' and 'outcome.'

Experiment 1

Method

Participants

A total of 217 naive participants (107 males, 105 females, and 5 who failed to report gender information), 17 to 24 years old, were recruited for the study from the SUNY-Binghamton subject pool. Their participation in the study was one way by which they could meet a course requirement. This and all subsequent experiments were approved by the SUNY-Binghamton Institutional Review Board.

Apparatus and stimuli

The experiment was conducted on windows PCs in individual cubicles at SUNY-Binghamton. The screen resolution was 1930 x 1080 pixels and the monitors were 53.34 cm wide. The experiment used a custom program written in Python using the Psychopy2 library (Peirce, 2007). The participants used a standard computer mouse to provide their responses. Participants sat with their faces approximately 60 cm from the screen.

Four sets of two stimuli (X, Y) were used for the cues: sets included capital letters (P, D), shapes (solid circle, solid square), Greek capital letters (β, Ω), and symbols (%, +). All stimuli were black and approximately one-fifth of the screen high (450 by 490 pixels).

Three sets of two stimuli (O1, O2) were used for the outcomes. They were taken from the International Affective Picture System (IAPS, Lang, Bradley, & Cuthbert, 2008). Each picture in the IAPS has a rating (on a 10-point scale) on each of two dimensions: pleasantness (i.e., valence) and arousal. Table 1 lists the three sets of pictures used here along with their IAPS identifier as well as arousal and valence scores. The pictures were originally 1024 x 768 pixels but were scaled down to match the size of the cues.

The stimuli appeared superimposed over a context. There were three different types of rectangular contexts outlined by their borders (which were black diagonal strips, open black dots [very small circles], or a black checkerboard pattern, each on a white field). The background of the contexts inside the borders were a solid color: light blue **(**RGB: 53, 188, 212**)**, bottle green (RGB: 22, 130, 24**)**, or bright yellow (RGB: 225, 190, 51**).** Each background color was yoked to a border pattern (i.e., stripes were always paired with the light blue background, dots were always paired with

the bottle green background, and checkerboards were always paired with the bright yellow background). Borders were the top 20%, right-most 20%, bottom 20%, and left-most 20% of the screen. More precisely, in terms of pixels, the borders were 1024 x 768 images, while the internal colored area was 960 x 662 pixels. A white fixation cross (10 x 40 pixels for the vertical line, 46 x 9 pixels for the horizontal line) was displayed at appropriate times at the center of the screen.

The above stimuli and contexts were used only during experimental training and testing. A special set of stimuli (key, table, tree, and car), appearing in a context with a red background (RGB: 232, 53, 57) and a border with a pattern of small Xs, was used during the warmup conditions that preceded the experimental conditions.

Procedure

Before participating in the experiment, all participants were required to complete an informed consent form and turn off their cell phones. Upon giving consent, they were led into individual experimental cubicles.

During the experiment, participants saw streams of trials after which they were asked to assess the conditional probability of one stimulus appearing given the appearance of the other stimulus (expectancy rating). There were three types of trial streams corresponding to the three conditions: Control (i.e., target training without any subsequent treatment intended to disrupt later test performance), CC, and Extinction. All participants experienced all three conditions (i.e., a within-subject design was used). For each participant, three sets of cues were picked randomly without replacement and assigned randomly without replacement to each of the three conditions (Control, Extinction, CC). In the same manner, a set of outcomes was randomly assigned without replacement to each of the three conditions.

During training, cue stimuli from one set (X, Y) and outcomes from the paired set (O1, O2) were presented to the participant. X and O1 were the target cue and outcome with respect to the expectancy rating. Y and O2 were alternative cues and outcomes. X was always presented one stimulus width to the right of the center of the screen, and Y was always presented one stimulus width to the left of center. Outcome O1 was always presented in the lower-left corner, diagonally

opposite X. Outcome O2 was always presented centered immediately below the fixation cross. The top panel of Figure 1 illustrates where on the screen each stimulus was presented on trials in which that stimulus was presented. For half of the participants (see below for details), the appetitive outcome was used as O1 and hence appeared in Phase 1 of a condition, whereas the aversive outcome was used as O2 and hence appeared in Phase 2 of a condition. For the remaining participants, O1 and O2 were reversed in affect.

Treatment for each experimental condition was composed of two phases separated by a 1000-ms grey screen during which only the fixation cross was visible. Phase 1 was used to establish an association between X and O1, whereas Phase 2 was used to disturb it through either CC or extinction (except for the Control condition in which the target association was intended to be left undisturbed apart from the passage of time and presentation of irrelevant stimuli). Each phase started with presentation of the context and the fixation cross for 2000 ms before the stream of trials started. During each trial, a cue appeared (X or Y), sometimes accompanied by an outcome (O1 in Phase 1 and O2 in Phase 2). Each trial was 400 ms long and followed by a 250-ms intertrial interval (ITI) during which only the fixation cross was present.

At the end of each treatment stream (i.e., after Phases 1 and 2), the following question appeared in the upper part of the colored background of the context: “Among all the trials on which X was presented, what was the percentage of trials on which O1 was also presented?” Instead of the names ‘X’ and ‘O1’ in the preceding sentence, the participants saw small representations of the stimuli (54 x 59 pixels) as they had seen them during training for that condition. The possible answers were presented below the question as a horizontal Likert scale of 0-100 with clickable circles at 0, 10, 20, … 90, and 100 (stretching across the lower fourth fifth of the screen in the middle three quadrants). The bottom panel of Figure 1 provides an illustration. The clickable circles had 0, 10, 20, etc. written inside of them and turned black for 150 ms when clicked upon. The scale was anchored at the left end by 'Never’ above the 0 and at the right end by 'Always' above the 100. The question remained on the screen until the participant responded. This expectancy rating was

followed by a 5-s screen with the words "Please wait" on it. Another screen then appeared which prompted participants to left-click the mouse when they were ready to start the next condition.

Warmup. Prior to the experimental conditions, all participants went through a series of warmup conditions aimed at familiarizing them with the procedure as well as providing them with examples of different conditional probabilities of O1 given X. The conditions during the warmup all used the specific context and stimulus set designated for it, which were distinct from those used in the experimental conditions.

Initially, an instruction screen welcomed participants, provided instructions to watch the screen, and prompted them to left-click the mouse to start the first warmup condition. In this warmup condition, O1 was always presented when X was presented (see Table 2). Phase 1 was composed of 6 trials in which X was presented simultaneously with O1 and 5 trials in which Y was presented alone: (100% training, Phase 1). During Phase 2, training consisted of 2 cycles of trials during each of which Y was presented simultaneously with O2 6 times and 6 more trials in which Y was presented alone (irrelevant training with respect to X and O1). That is, the program cycled once through Phase 1 and then twice through Phase 2. During each cycle, the order of presentation of the trials was determined randomly. Once the participants provided their response on the Likert scale, an instruction screen appeared explaining to the participants that, as cue X was paired with outcome O1 on 100% of the trials on which X was presented, their answer should have been 100%. Participants were then prompted to left-click the mouse to start the next condition.

In the second warmup condition, X and O1 were never presented together (0% training, see Table 2). Phase 1 was composed of 6 trials on which X was presented by itself and 5 trials on which Y was presented with O2. Phase 2 was identical to Phase 2 for the 100%- warmup condition. As was done with the positive warmup condition, Phase 1 was presented once whereas Phase 2 was cycled through twice, with the order of presentation of the trials determined randomly during each cycle. Once the participant provided their response on the Likert scale, an instruction screen appeared explaining to the participant that, as cue X was paired with outcome O1 on none of the trials on

which it was presented, the answer should have been 0%. The participant was then prompted to left-click the mouse to start the next warmup condition.

In the final warmup condition, O1 was shown only on half of the trial in which X was presented (50% training, see Table 2). The single cycle through Phase 1 and both cycles through Phase 2 were each composed of 6 trials in which X was presented, 3 of which were with O1, and 5 trials in which Y was presented, 3 of which were with O2. Once the participant provided a response on the Likert scale, a screen appeared explaining to the participant that because, cue X was paired with outcome O1 on half of the trials on which it was presented, the answer should have been 50%.

Participants were then told they would be presented with more examples of each of the three types of conditions and that they should try to identify them as accurately as possible. Upon left-clicking the mouse, participants were exposed to series of condition triplets consisting of the 100% training, 0% training, and 50% training conditions (each once) in a random order for each triplet. Following each warmup training condition, the participant's response was defined as 'correct' if it was 80%, 90%, or 100% for a 100% warmup condition, 30%, 40%, 50%, 60%, or 70% for a 50% warmup condition, and 0%, 10%, or 20%, for a 0% warmup condition. Participants were repeatedly presented with this triplet of warmup conditions until they were either able to provide a 'correct' response for each condition composing a triplet for two triplets in a row or until they had been exposed to 10 triplets. In the latter case, they were considered to have failed warmup training, and were thanked and dismissed. Thirty-five participants failed to meet the learning criterion during the warmup conditions.

Experimental conditions. The experimental conditions started immediately after the warmup criterion was reached. An instruction screen informed participants that the relation between X and O1 would now be more difficult to detect and prompted them to click the mouse to start the study. They were then exposed to 10 triplets of conditions, each triplet composed of a control condition, an extinction condition, and a CC condition (see Table 2). This allowed for a better assessment of the outcome expectancy than if a condition was presented only once to the participant. As indicated previously, stimuli were assigned randomly to each of the three conditions. Once assigned to a

condition, they remained so through the entire study. Phase 1 was identical across all three conditions and was designed to establish an association between X and O1. It consisted of 6 presentations of X, 5 of which were accompanied by O1, 1 of X alone, and 5 presentations of Y alone. The order of presentation of these 11 trials was determined randomly. For the Control condition, Phase 2 consisted of two cycles each of 11 presentations of Y, 5 of them with O2 and 6 of them without an outcome, presumably leaving the X-O1 association established in Phase 1 unaffected. In contrast, in the extinction and CC conditions, the Phase 2 trials were designed to disrupt responding based on the X-O1 association that was established in Phase 1. In the Extinction condition, each of two cycles through Phase 2 consisted of 5 presentations of X alone, and 5 presentations of Y-O2 plus 1 presentation of Y alone. For the CC condition, each of two cycles through Phase 2 consisted of 5 presentations of X with O2 and 6 presentations of Y alone. The order of presentation of the trials within each of the two cycles of Phase 2 trials was randomly determined for each participant. After providing a response to the test question, the participant was prompted to start the next condition by left clicking the mouse.

The order of presentation of the three conditions during the first triplet after the warmup conditions was counterbalanced across participants. Due to a programming error, the order in the subsequent triplets remained identical to the one in the first triplet, instead of being determined randomly as originally intended (this programming error carried over to all the other experiments reported in the present article); consequently, the order of the three types of conditions was fully counterbalanced across participants within a triplet, but was unvaried across the 10 repetitions through the three conditions for each participant. Once the participant had completed a triplet, a 2000-ms grey screen was presented before the start of the next triplet, which asked the participant to left-click the mouse when ready.

Assessing valence of outcomes. After presentation and rating of all conditions ten times, the affective valences of the cues and outcomes were assessed. Participants were asked to evaluate the valence of each cue and each outcome to which they had been exposed on a scale of -5 to +5 scale. To obtain a rating, each stimulus was shown for 400 ms centered on the screen over a black

background. This was followed by the question *"How pleasant or unpleasant is this image for you?"* along with the presentation, below the question, of an 11-point Likert scale going from -5 to +5 and anchored at -5 (very unpleasant), 0 (neither pleasant nor unpleasant) and +5 (very pleasant). Once the participant provided a response by clicking on the Likert scale, a 1000-ms intertrial screen, consisting of a white fixation cross over a dark background, was shown. The next stimulus was then presented. The 6 cues (X and Y for each of the three conditions) were shown in a random order, 5 times each. Then the 6 outcomes (O1 and O2 for each of the three conditions) were shown in a random order, 5 times each. The cues were shown first to preserve sensitivity, if any, in case of carryover effects from the assessment of the IAPS images that were used as outcomes. Once the valence ratings were completed, participants were presented with a debriefing screen informing them of the intent of the experiment.

Data analysis

Though a Likert scale is formally an ordinal scale, we treated it, as is often the case in the literature, as an interval scale (see Maia et al, 2018, for a critic of that approach). For each participant and each condition (Control, Extinction, and CC), a mean US expectancy rating was computed based on the 10 ratings per condition that the participant had provided. Likewise, for each stimulus presented during the valence rating phase, a mean rating was computed based on the 5 ratings the participant gave for each stimulus.

Inferential analyses on these dependent variables of expectancy and valence rating were carried out based on 95% confidence intervals (CI) computed using Student's t-distribution. Cohen's *d* was used as a measure of effect size for all pairwise comparisons (Control minus Ext, Control minus CC, CC minus Ext). Following Cummings' (2012) recommendation for the computation of Cohen's *d* in a within-subject design, it was computed using the formula

$$d = \left[1 - \frac{3}{4(n-1)-1}\right] \frac{\sum_{i=1}^{n}(x_{i1} - x_{i2})}{n\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

where x_{ij} is the score of participant *i* in condition *j,* and *n* is the number of participants. 95% CI for Cohen's *d* was computed using ESCI (https://thenewstatistics.com/itns/esci/) using the method described by Algina & Kesselman (2003). This method provides a reasonable approximation of the 95% CI for Cohen's *d* in a paired design if (a) the number of participants is larger than 10, (b) Cohen's *d* in the population is between -1.8 and 1.8, and (c) the correlation in the population between the two conditions is between 0 and 0.8. If, for some reason, these conditions are not met (for instance, if the observed Cohen's *d* is over 1.8, which would in any case indicate a very large effect), it is not possible to compute 95% CI with this method. To the best of our knowledge, no alternative method exists at this time.

Besides the 35 participants who failed to meet the warmup criterion during the warmup conditions, four additional participants failed to complete the experimental part of the study. Critically, because of the possibility of gender-related differences in the way males and females process affectively-loaded stimuli, we decided that the number of female participants for whom O1 was appetitive should approximate the number of male participants for whom O1 was appetitive, and similarly when O1 was aversive. As we could not foresee which participants would fail the warmup criterion or other problems (e.g., one participant was removed from the analysis for failing to specify gender) which would not allow us to include them in the analysis, a surplus of participants was initially included. These surplus participants were not included in the analysis because only the data from the participants required to complete the counterbalancing were retained. Based on the lowest number of remaining participants of either gender for whom O1 was of one specific valence, we eliminated participants using the rule 'last run, first eliminated,' until we had a balance of genders between the two valences of O1. This left 142 participants. For 71 of them (36 males and 35 females), O1 was appetitive, whereas for the remaining 71 (36 males and 35 females), O1 was aversive. Due to the within-subject nature of the 3 conditions (Control, CC, and Extinction), this automatically counterbalanced gender with respect to condition. Counterbalancing of gender across affective value of outcomes was also approximated in Experiment 2 because O1s were affectively-

loaded stimuli in that experiment. No effort was made to counterbalance gender in Experiments 3 and 4 because the outcomes were of neutral affect.

Although the statistical analyses were performed using confidence intervals instead of null hypothesis testing, we can still examine how much statistical power a t-test would have if we had used one on our data. With a sample size of 142, assuming a worst-case scenario in which performance in one condition is not correlated with performance in another (making our paired-group design similar to an independent group one), the statistical power for a t-test is 99% for a Cohen's *d* of 0.5, 92% for a Cohen's *d* of 0.4, and 71% for a Cohen's *d* of 0.3. This level of statistical power is achieved only if we do not take into account the valences of O1 and O2. Otherwise, if we look at the difference between CC and extinction only for the participants for whom O1 was appetitive (and hence O2 aversive), the sample size is only 71; hence, power decreases to 84% for a Cohen's *d* of 0.5, 66% for a Cohen's *d* of 0.4, and 43% for a Cohen's *d* of 0.3. Given those probabilities, the analyses were carried out only at the level of the 142 participants. Analysis of gender failed to detect any consistent differences. Hence, the analyses of gender for this experiment and the subsequent ones are reported only in the appendix.

Results

Outcome valence

Figure 2 shows the result of the valence ratings for the outcomes obtained at the end of Experiment 1 as a function of the type of streams. The appetitive outcomes were all rated positive on the Likert scale while the aversive outcomes were all rated negative, showing that the fast stimulus presentation and the reduced size of the pictures did not appreciably affect their emotional impact. Also, there were no appreciable differences between conditions (i.e., Control vs. Extinction vs. CC) in terms of outcome valence.

Evaluative conditioning

In the Control condition, if the emotional valence of the cue X is affected by its pairing with the outcome O1 (evaluative conditioning), then, when participants are asked about the valence of X at the end of the study, they should report a positive rating if O1 was appetitive and a negative rating

if O1 was aversive. Our sample size was too small to meaningfully examine the valence of X separately when O1 was appetitive and when it was aversive. Hence, the data from all 142 participants were pooled in the following manner: the ratings given by the participants for whom O1 was appetitive were left unchanged, and the ratings given by the participants for whom O1 was aversive were multiplied by -1. This way, independent of the valence of O1, a positive rating indicated a rating consistent with evaluative conditioning.

The means of the modified Likert ratings are shown in the top panel of Figure 3. Based on this figure, there is some evidence of evaluative conditioning in the Control and Extinction conditions. In contrast, the bottom panel of Figure 3 suggests that at the descriptive level, CC was more efficient at reducing evaluative conditioning than was extinction, but the data did not support this conclusion at the inferential level (Control vs. Extinction: 0.08, 95% CI [-0.05, 0.21], Cohen's *d* = 0.08, 95% CI [-0.05, 0.20]; Control vs. CC: 0.19, 95% CI [-0.04; 0.42], Cohen's *d* = 0.17, 95% CI [-0.03, 0.372]; Extinction vs. CC: 0.11, 95% CI [-0.09, 0.31], Cohen's *d* = 0.10, 95% CI [-0.09, 0.29]).

Outcome expectancy

Figure 4 shows the mean expectancy ratings for O1 in the presence of X in Experiment 1 as a function of condition (top panel) as well as the mean differences in Likert ratings (bottom panel). Both extinction and CC successfully decreased the assessment of the likelihood of O1 in the presence of X (mean difference in Likert rating: Control vs. Extinction: 34.85, 95% CI [31.82, 37.88], Cohen's *d* = 2.33; Control vs. CC: 31.95, 95% CI [28.83, 36.07], Cohen's *d* = 2.12), but CC was less efficient than extinction at doing so (mean difference in Likert rating: Extinction vs. CC: -2.90, 95% CI [-4.44, -1.36], Cohen's *d* = -0.16, 95% CI [-0.24, -0.07]).

Discussion

The present high-powered study with 142 participants aimed to detect differences between CC and extinction. We attempted to detect a difference between the two procedures on a variable reflecting evaluative conditioning (ratings of the valence of the target cue), on an emotional type of

processing, and on ratings of the O1 expectancy in presence of X, a more cognitive type of processing.

Although not statistically significant, the evaluative conditioning data are consistent with a greater efficiency of CC than extinction at reducing evaluative conditioning, a result already reported by others (Dunsmoor et al., 2019; Engelhard et al. 2018; Kerkhof et al, 2011; Reynolds et al, 2018). The failure to detect an effect despite the large number of participants indicates that, if this effect exists at all, it is quite small: the 95% CI for Cohen's *d* between extinction and CC ranges from -0.09 to 0.29. Thus, to achieve at least 80% power with a t-test in an independent-group design, one would need to run more than 1000 participants if the population effect size is on the lower end of the CI and more than 175 if it is on the upper hand of the CI. This is much more than the median sample size in the studies which have looked at the differential effect of CC and extinction on evaluative conditioning. Some of these studies have used implicit measures of evaluative conditioning such as emotional priming (Enghelardt et al., 2014; Kerkhof et al., 2011) or physiological measures (Dunsmoor et al. 2019; Reynolds et al., 2018). These implicit measures might be more sensitive to differences between CC and extinction than the explicit measures used here. But at least some of these studies (for instance, Kerkhof et al., 2011) have reported a similar effect with verbal measures very similar to the one used here, so this cannot be the full basis of the discrepancy.

In contrast, a clear but small difference was observed between extinction and CC in the O1 expectancy in the presence of X, although the effect was not the one expected. Based on the nonhuman animal data, we expected to find lower O1 expectancy ratings following CC treatment. Instead, CC was less efficient than extinction at reducing O1 expectancy. One possible factor explaining this surprising result is the unusually short duration of the stimuli we used. In order to determine whether this was a factor, Experiment 2 replicated Experiment 1 but with a longer stimulus duration.

Experiment 2

Method

Participants and apparatus

A total of 201 naive participants (71 males, 110 females, and 20 who failed to report gender information), 18 to 23 years old, were recruited for the study from the SUNY-Binghamton subject pool. Apparatus and stimuli were identical to the ones used in Experiment 1.

Procedure

The procedure was identical to the one used in Experiment 1, except for the following modifications. During both the warmup and experimental parts of the study, the cues and outcomes were presented for 1250 ms instead of 400 ms, and the ITI was 1250 ms instead of 250 ms. During the valence rating phase, stimuli were also presented for 1250 ms instead of 400 ms.

As the increased duration of the stimuli and ITI lengthened the duration of the experimental session, the following additional changes were also made. During the warmup, the program considered that participants failed training if they did not provide the expected responses to two triplets of the warmup conditions (i.e., 100%, 0%, and 50%) in a row after having been exposed to 5 triplets (as opposed to 10 in Experiment 1). This change in the warmup criterion did not appreciably affect the rate of participants passing the warmup criterion as only 28 participants failed to meet the criterion during the warmup conditions. During the experimental part of the study, participants were exposed to triplets of experimental conditions (i.e., Control, Extinction, and CC) 3 times (as opposed to 10 times Experiment 1).

Data analysis

The same counterbalancing used in Experiment 1 was used in Experiment 2. One hundred forty-four participants were selected from the 173 remaining participants: The 29 participants excluded from the analysis were the ones run in surplus (i.e., last) in order to achieve optimal counterbalancing. Despite this effort, the counterbalancing was not perfect. For half of the 144 participants (32 males, 36 females, 4 participants failed to provide gender information, 18 to 21 years old), O1 was appetitive, while for the remaining half (28 males, 38 females, 6 participants failed to provide gender information, 18 to 23 years old) O1 was aversive. Otherwise, the data were analysed in the same manner as in Experiment 1.

Results

Outcome valence

Figure 5 shows the mean Likert rating for the valence of Experiment 2 outcomes as a function of condition. As in Experiment 1, all the appetitive outcomes were rated positively while all the aversive outcomes were rated negatively, with no obvious difference between conditions. This confirms our presuppositions concerning the valence of the outcomes.

Evaluative conditioning

As in Experiment 1, the ratings for the participants for whom X was paired with an aversive outcome in Phase 1 were multiplied by -1 so that a positive rating was always consistent with evaluative conditioning. The top panel of Figure 6 depicts the mean Likert rating for the valence of cue X in Experiment 2 as a function of condition, whereas the bottom panel of Figure 6 illustrates the comparisons between conditions. The data are highly similar to those obtained in Experiment 1. At the descriptive level, they point to a potential reduction of evaluative conditioning with both extinction and CC, though CC seems to be somewhat more efficient. This time, at the inferential level, the claim that the ratings are lower in the CC condition than in the Control condition is supported, while the claim that ratings are higher in the Extinction condition compared to the CC conditionis not supported. The effect sizes here, whether they are supported at the inferential level or not, remain quite small (mean difference in Likert rating: Control vs. Extinction: 0.13, 95% CI [-0.07, 0.32], Cohen's $d = 0.10$, 95% CI [-0.05, 0.25]; Control vs. CC: 0.33, 95% CI [0.09, 0.56], Cohen's $d = 0.26$, 95% CI [0.07, 0.44]; Extinction vs. CC: 0.20, 95% CI [-0.08; 0.48], Cohen's $d = 0.15$, 95% CI [-0.06; 0.22]).

Outcome expectancy

Figure 7 shows the mean Likert ratings for the O1 expectancy in the presence of X in Experiment 2 as a function of condition (top panel) as well as the mean differences in Likert ratings (bottom panel). The results are nearly identical to those found in Experiment 1: relative to the Control condition the O1 expectancy ratings were much lower after extinction (mean difference in Likert rating for Control vs. Extinction conditions: 40.16, 95% CI [37.41, 42.92], Cohen's $d = 2.97$)

and CC (mean difference in Likert rating for Control vs. CC conditions: 36.50, 95% CI [33.64, 39.36], Cohen's *d* = 2.67) and, once more, extinction proved itself more efficient than CC (mean difference in Likert rating for Extinction vs. CC conditions: -3.66, 95% CI [-6.15, -1.16], Cohen's *d* = -0.29, 95% CI [-0.49, -0.09]).

Discussion

Overall, the data from Experiment 2 are in line with those from Experiment 1 and demonstrate that the short stimulus durations used in Experiment 1 were not a factor in the results. Looking at the efficiency of extinction and CC at reducing evaluative conditioning, the pattern is strikingly similar to the one observed in Experiment 1. That is, at the descriptive level, it seems that CC is more efficient than extinction. However, the effects were very small, and, despite our large sample size, we still did not have strong support for this conclusion at the inferential level. Yet, as the stimulus duration does not seem to matter, Experiment 2 can be considered a replication of Experiment 1 and, following the method described by Cummings (2012), the two experiments can be combined in a meta-analysis to obtain a better estimate of effect size. The meta-analysis was performed using ESCI (https://thenewstatistics.com/itns/esci/, see Cummings, 2012 for further details). It provides support for the claim that CC reduces evaluative conditioning (Cohen's *d* for the Control vs. CC comparison: 0.24, 95% CI [0.08, 0.39]) and that is more efficient at doing so than is extinction (Cohen's *d* for the Extinction vs. CC comparison: 0.14, 95% CI [0.02, 0.27]). In contrast, there is still no support for the claim that extinction interferes with evaluative conditioning (Cohen's *d* for the Control vs. Extinction comparison: 0.10, 95% CI [-0.02, 0.22]). These results are illustrated in Figure 8. This is in line with other reports in the literature (Dunsmoor et al., 2019; Engelhard et al. 2018; Kerkhof et al, 2011; Reynolds et al, 2018), though the effects here are much smaller than the ones reported in those papers.

The O1 expectancy judgment results of Experiment 2 also replicated those of Experiment 1. Once more, CC was found to be less efficient than extinction at reducing the O1 expectancy in the presence of X. A meta-analysis performed on both experiments returns a Cohen's *d* for the Extinction vs. CC comparison of 0.22, 95% CI [0.10, 0.35].

As we now had good reason to think that CC was more efficient at reducing evaluative conditioning, the discrepancy between our measure of evaluative conditioning and our measure of the outcome expectancy raised an interesting issue. There are two possible ways to explain a difference in the efficiency of CC relative to the one of extinction. The first is in terms of emotional factors. Extinction and CC would both be effective because they allow the cue to trigger an emotional state counteracting the emotional state initially triggered by the cue following conditioning, but CC produces a greater change in emotional state than does extinction. The second would emphasize cognitive factors, by emphasizing the similarity between extinction and CC as retroactive interference paradigms used to produce forgetting in memory research (e.g., Polack, Jozefowiez, & Miller, 2017). It is reasonable to assume that the valence ratings tap into emotional processes, while the outcome expectancy measure depends more exclusively on memory processes. Hence, the conclusion to draw from Experiments 1 and 2 would be that CC is more efficient at interfering with emotional processes but less efficient at interfering with memory processes. If this account is correct, extinction should still be more efficient than CC even if neutral outcomes were used. This prediction was tested in Experiment 3, which replicated Experiment 1 but with neutral outcomes.

Experiment 3

Method

Participants

A total of 100 naive participants (30 males, 61 females, and 9 who failed to report gender information), 17 to 36 years old, were recruited for the study. Eighty-eight of them (17 to 36 years old, 22 males and 57 females, with 9 participants failing to report gender information) were recruited from the SUNY-Binghamton subject pool, whereas the remaining 12 (18 to 24 years old, 8 males and 4 females) volunteered at the University of Lille.

Apparatus and stimuli

The computers in Binghamton were identical to the ones used in Experiment 1. The computers in Lille had a screen resolution of 1280 x 1024 and the monitors were 34 cm wide. The

program adjusted the size of all the stimuli to the size of the monitors so that the size of the borders on the x-axis corresponded to the size of the monitor on the x-axis. Otherwise, the material was identical to the one used in Experiment 1 except for the cues and outcomes which were now organized into four sets of four stimuli (X, Y, O1, O2): Capital letters (P, D, M, Z), shapes (solid circle, solid square, solid triangle, and solid diagonally-oriented disk), Greek capital letters (β, Ω, π, and ψ), and symbols (%, +, ^, *). All stimuli were black and approximately one-fifth of the screen high (450 by 490 pixels). As in the prior experiments, the quartet of stimuli assigned to each condition was randomly determined for each participant, and the role of each stimulus within a quartet was randomly assigned for each participant. Note that the cues were identical to the ones used in Experiments 1 and 2.

Procedure and data analysis

The procedure was identical to the one used in Experiment 1. The only differences were the use of the neutral outcomes O1 and O2 instead of affectively-loaded ones, and no valence rating phase occurred at the end of the experiment. The data were analyzed as in Experiment 1.

Results

The top panel of Figure 9 shows the mean O1 expectancy ratings in Experiment 3 as a function of condition, while the bottom panel shows the difference in the Likert rating for each pair of conditions. The results are essentially identical to the ones observed in Experiments 1 and 2. First, extinction and CC decreased the assessment of the outcome expectancy in the presence of the cue (mean difference between the Control and Extinction conditions: 37.20, 95% CI [33.37, 41.03], Cohen's d = 2.75; mean difference between the Control and CC conditions: 32.30, 95% CI [28.54, 36.06], Cohen's d = 2.30). Second, extinction once again proved to be somewhat more efficient at lowering the expectancy judgment than CC (mean difference between Extinction and CC: -4.90, 95% CI [-7.25, -2.54], Cohen's d = -0.34, 95% CI [-0.52, -0.17]).

Discussion

The results of Experiment 3 replicate those of Experiments 1 and 2 despite the outcomes used in Experiment 3 lacking any appreciable emotional valence. It provides further confirmation that, at

least in this preparation, extinction is more efficient than CC at decreasing expectancy judgments and confirms that it does so without involving any kind of emotional processing. That is, the reduction presumably acts through a more cognitive route potentially involving interference in memory.

A pattern started to emerge here. Overall, CC seems more efficient than extinction at altering emotional processing, a result corroborated by other studies of evaluative conditioning in humans (De Jong et al., 2000; Dunsmoor et al., 2019; Engelhard et al., 2018; Kang et al., 2018; Kerkhof et al., 2011; Meulders et al, 2015; Raes & De Raedt, 2012; Reynolds et al, 2018). In contrast, extinction is more efficient than CC at reducing outcome expectancy in the presence of the cue, a measure that might reflect processing in a more cognitive domain than evaluative conditioning. Because this effect is quite small (if the meta-analysis is updated to take into account Experiment 3, Cohen's *d* is now equal to 0.25, 95% CI [0.15, 0.36]), it is not surprising that it had not been detected by prior studies because they lacked sufficient statistical power.

We are, however, still left with the question of why nonhuman animal research (Escobar et al., 2001; Holmes et al., 2016; Tunstall et al., 2012) has found CC to be more efficient than extinction. One possibility is that the animal studies tapped into emotional processing, but Escobar et al.'s (2001) demonstration used a sensory preconditioning paradigm in which both Phase 1 target training and Phase 2 decremental training was conducted with stimuli that were affectively nearly neutral, and consequently is, in some ways, similar to the present Experiment 3. Another possibility is that, because the outcome expectancy question used in the present research mentions O1, it is somehow similar to a reinstatement procedure in which the subject is re-exposed to US1 being presented soon before or at test. This is not the way testing is usually done in animal studies; there the subject is simply re-exposed to the CS following extinction or CC. As O1 only appears in Phase 1 and, if for some reason, CC is more susceptible to reinstatement than extinction, it would explain why we consistently found extinction to be more efficient than CC in our preparation. Following CC, the outcome expectancy question would prompt the participant to selectively retrieve the phase in which O1 was presented, in this case Phase 1, in which X and O1 were paired, thereby decreasing

the efficiency of CC. If this explanation is correct, then the difference between CC and extinction should be eliminated if O1 is also presented during Phase 2 of treatment. This prediction was tested in Experiment 4.

Experiment 4

Method

Participants and apparatus

A total of 87 naive participants (29 males, 48 females, and 10 who failed to report gender information), 17 to 25 years old, were recruited from the SUNY-Binghamton subject pool. The apparatus and some of the stimuli were identical to those used in Experiment 1. To replicate Experiment 3 and to examine the consequences of adding O1-alone exposures to Phase 2, six new sets of cue-outcome stimuli (X, Y, O1, O2) were constituted: set A (basketball, lightbulb, downward-facing triangle, ^), set B (pear, building, octagon, arch), set C (hammer, keyboard, square, Z), set D (flower, umbrella, P, star), set E (book, plane, omega, leftward arrow), and set F (computer mouse, tie, percentage sign, oval). Three new sets of context backgrounds were also added to those used in Experiments 1 to 3: orange background (RGB: 253, 150, 73) over a star pattern; pink background (RGB: 253, 100, 149) over a wavelet pattern; and purple background (RGB: 196, 107, 252), all over a pattern composed of repeated 8-shaped forms. The context and stimuli used during treatment were then configured in the same manner as in Experiments 1 to 3.

Procedure and data analysis

The procedure was identical to that used in Experiment 3, except for the following modifications. During warmup, 5 O1-alone trials were randomly interspersed among the other trials during each cycle of Phase 2 in all three warmup conditions (see Table 4). During those trials, only outcome O1 was presented on the screen.

During the experimental part of the Experiment, three new types of conditions were added. Each of these conditions was identical to one of the original conditions, except that during each cycle of Phase 2, 5 O1-alone trials were interspersed among the other trials (see Table 5). Thus, for instance, treatment in the Control-O1 condition was identical to that of the Control condition, except

for the addition of 5 O1-alone trials during each of the two cycles of Phase 2. Which of these six conditions a participant experienced first was counterbalanced across participants. The order of presentation of the remaining five conditions was determined randomly for each participant. A participant was exposed to each of the six conditions in each of the five blocks. For a given participant, a set of cue-outcome stimuli and context pictures was assigned randomly without replacement to each of the 6 conditions and maintained throughout all five repetitions of that condition.

As some of the stimuli that were used were pictures of meaningful objects, their valence was assessed to make sure they were close to neutral in affect. Hence, the experimental conditions were followed by an affective rating phase identical to the one used in Experiments 1 and 2. Stimuli were shown for 400 ms as in training. Due to a programming error, participants rated only 3 sets of cues and outcomes instead of the 6 to which they had been exposed (the 3 sets were the ones used for the 3 first conditions to which they had been exposed, which were randomized across subjects). The data analysis followed the procedures used in Experiment 1.

Results

Thirty-three participants failed to meet the warmup criterion and 1 participant failed to complete the experimental part of the study. The analysis was based on the remaining 53 participants (19 males, 29 females, and 5 participants who failed to provide gender information).

Cue and outcome valence

Figure 10 shows the mean Likert scores for the valence rating of the cues and outcomes in Experiment 4. Overall, the cues are rated a bit higher than the ones in Figures 2 and 4, which could reflect either that on average the ones used as outcomes were more meaningful in Experiment 4 than the cues in Experiments 1 and 2, or that the lack of strongly affective outcomes caused the participants to rate the cues higher. The outcomes appeared to be rated slightly lower than the cues, which is consistent with the fact that we used less meaningful stimuli for them. However, Figure 10 confirms that all stimuli, especially the outcomes, had little emotional content.

Outcome expectancy

Figure 11 shows the mean Likert ratings for the O1 expectancy in the presence of X (top) and mean differences in Likert ratings (bottom) in Experiment 4 as a function of condition. For the conditions in which O1 was presented only during Phase 1, the results of Experiments 1 to 3 were essentially replicated: Extinction and CC both interfered with the outcome expectancy relative to the Control condition (mean difference in Likert rating: Control vs. Extinction, 35.81, 95% CI [30.29; 41.33], Cohen's d = 2.41; Control vs. CC, 28.11, 95% CI [21.32, 34.90], Cohen's d = 1.73, 95% CI [1.20, 2.25]) and Extinction was more efficient at doing so than CC (mean difference in Likert rating between extinction and CC: -7.70, 95% CI [-11.75, -3.64], Cohen's d = -0.54, 95% CI [-0.84, -0.24]).

For the conditions in which O1 was presented during both Phase 1 and Phase 2, the outcome expectancy was also decreased by extinction (mean difference in Likert rating between Control and Extinction: 32.38, 95% CI [26.74, 38.01], Cohen's d = 2.24) and CC (mean difference in Likert rating between Control and Extinction: 36.68, 95% CI [30.68, 42.67], Cohen's d = 2.53), but CC now proved more efficient than Extinction at doing so (mean difference in Likert rating between extinction and CC: 4.30, 95% CI [0.73, 7.87], Cohen's d = 0.34, 95% CI [0.055, 0.62]). Indeed, as suggested by the top panel of Figure 11 and confirmed by Figure 12, the introduction of O1 during Phase 2 appears to have solely impacted the effect of the CC treatment (mean difference in Likert rating for Control vs. Control-O1: 2.83, 95% CI [-2.01, 7.68], Cohen's d = 0.17, 95% CI [-0.12, 0.47]; Extinction vs. Extinction-O1: -0.60, 95% CI [-4.31, 3.11], Cohen's d = -0.05, 95% CI [-0.343, 0.245]; CC vs. CC-O1: 11.40, 95% CI [5.40, 16.79], Cohen's d = 0.79, 95% CI [0.39, 1.18]).

Discussion

The first conclusion from Experiment 4 is that it further confirms the results of Experiments 1-3 concerning the greater efficiency of extinction over CC in reducing outcome expectancy in the present preparation. If the results of the meta-analysis are updated to include Experiment 4, Cohen's d for this effect is now equal to 0.30, 95% CI [0.17, 0.43]. Experiment 4 suggests that this is the result of a sort of reinstatement induced by the mention of O1 in the question aimed at assessing the

outcome expectancy (i.e., an ABA renewal procedure): CC becomes more efficient than extinction at reducing the O1 expectancy in the presence of X if O1 is presented in Phase 2 to preclude selective reinstatement of Phase 1 (i.e., creating an AAA renewal control procedure). This could explain the discrepancy between our results and the nonhuman animal data (Escobar et al., 2001; Holmes et al., 2016; Tunstall et al., 2012), though the effect size remains much smaller in our case. One implication is that, if nonhuman animals were re-exposed to the target US (O1) soon before testing with the CS, CC might prove less efficient than extinction. Overall, the present data suggest that, beyond reinstatement per se, CC might be more susceptible than extinction to recovery effects, as suggested by Holmes et al. This is a question worth pursuing in future research. Previously published data to date have been ambiguous. In rats, Holmes et al. reported a greater sensitivity of CC to renewal, while Dunsmoor et al. (2015) reached the opposite conclusion using a spontaneous recovery design. With human participants, both Kang et al. (2018 assessing renewal) and Dunsmoor et al. (2015, 2019 assessing spontaneous recovery) have concluded that CC was less susceptible to recovery effects than extinction.

Experiment 5

We consistently observed differences between extinction and counterconditioning in all the experiments reported above, although these differences were small: Collectively, they yielded a Cohen's *d* of 0.3 in the cognitive domain (i.e., outcome expectancy ratings), and a Cohen's *d* of 0.2 in the emotional domain. A potential explanation for this is that the participants have difficulty grasping the X-O2 relation in the CC condition. As a consequence, they would process only a subset of the X-O2 trials as such. The rest would be processed as nonreinforced X Trials. If so, they would be expected to process the CC condition almost like the Extinction one. This would account readily for the small difference in the expectancy rating between the two. If this were the case, it would undermine the conclusions we drew from the previous experiments. As we never queried the participants about the X-O2 relation, we cannot rule out this hypothesis. Therefore, the goal of Experiment 5 was to assess the possibility that the similarity in the consequences of CC and

extinction in the prior experiments arose from impaired perception of the X-O2 pairings. Basically, it procedurally replicated Experiment 4, except that at the end of each condition the participants rated the O2 expectancy in the presence of X instead of the O1 expectancy in the presence of X. If the participants are processing most X-O2 trials as nonreinforced X trials (i.e., extinction trials), we should observe a very low O2 expectancy in the presence of X in the CC condition.

Method

Participants and apparatus

A total of 115 naive participants (64 males, 49 females, and 2 who failed to report gender information), 17 to 23 years old, were recruited for the study from the SUNY-Binghamton subject pool. The apparatus and the stimuli were identical to those used in Experiment 4.

Procedure and data analysis

The procedure was identical to the one used in Experiment 4, except that the participants were asked to rate how likely O2 was to appear in the presence of X instead of how likely O1 was in the presence of X. In addition, the participants were not asked to rate the valence of the cues and outcomes at the end of the study. Data analysis followed the protocol used in previous experiments.

Results

A total of 43 participants failed to meet the warmup criterion. The analysis was based on the remaining 72 participants (46 males, 25 females, and 1 participant who failed to provide gender information).

As Figure 13 shows, whether O1 was shown only in Phase 1 or not, the participants judged that X and O2 were strongly linked in the CC condition but not in either the Control or the Extinction conditions (mean difference in Likert ratings for Control vs. CC: O1 in Phase 1 only, -62.73, 95% CI [-68.07, -57.39], Cohen's d = -3.72, correlation = 0.07; O1 in Phase 1 and 2, -60.74, 95% CI [-66.96, -54.52], Cohen's d = -3.49, correlation = -0.18; mean different in Likert rating for Extinction vs. CC: O1 in Phase 1 only, -60.04, 95% CI [-65.42, -54.67], Cohen's d = -3.54, correlation = 0.07; O1 in Phase 1 and 2, -58.42, 95% CI [-64.15, -52.69], Cohen's d = -3.46, correlation = -0.07). However, there was no obvious difference between the Control and Extinction conditions (mean difference in

expectancy ratings for Control vs. Extinction: O1 in Phase 1, -2.68, 95% CI [-6.24, 0.87], Cohen's *d* = -0.15, 95% CI [-0.35, 0.05], correlation = -0.64; O1 in Phase 1 and 2, -2.31, 95% CI [-6.38, 1.75], Cohen's *d* = -0.14, 95% CI [-0.49, -0.14], correlation = 0.41]. As seen in Figure 14, although only at the descriptive level, the introduction of O1 in Phase 2 seems to have slightly negatively impacted the O2 expectancy in the presence of X. However, this conclusion was not supported at the level of statistical inference (mean difference in Likert rating: Control vs. Control-O1, -1.53, 95% CI [-5.33, 2.27], Cohen's *d* = -0.09, 95% CI [-0.31, 0.13], correlation = 0.55; Extinction vs. Extinction-O1, -1.89, 95% CI [--5.02, 1.22], Cohen's *d* = -0.11, 95% CI [-0.30, 0.07], correlation = 0.69; CC vs. CC-O1, -3.51, 95% CI [-7.50, -0.46], Cohen's *d* = 0.21, 95% CI [-0.44, 0.03], correlation = 0.50).

Discussion

Experiment 5 established that participants clearly differentiated between the Extinction and the CC condition: they clearly perceive that X and O2 are not linked in the former case whereas they are in the latter case. It provides a compelling counterargument to any concerns that the small differences observed between the Extinction and CC conditions in Experiments 1-4 were due merely to the participants failing to perceive the X-O2 relation in the CC condition.

The data from Experiment 5 also rule out another potential explanation for the results of Experiment 4. The greater efficiency of CC compared to Extinction in decreasing the O1 expectancy in the presence of X observed in Experiment 4 (when O1 was presented in both Phase 1 and Phase 2) could have been due to introduction of O1 in Phase 2 boosting the X-O2 association (or its retrievability at test), thereby making retroactive interference more potent. However, as Figure 14 shows, the data are not compatible with such a hypothesis. The CI for the effect size with respect to the impact of the introduction of O1 in Phase 2 in the CC condition suggests that this manipulation either had no meaningful effect (the upper bound for Cohen's *d* is 0.03) or led to a reduction of the O2 expectancy in the presence of X (the lower bound for Cohen's *d* was -0.44). Meaningful positive values for Cohen's *d* are ruled out by the data, which refutes the hypothesis that the introduction of O1 in Phase 2 boosted the X-O2 association, and hence, the O2 expectancy in the presence of X. At

best, introducing O1 in Phase 2 led to a weaker X-O2 association. Therefore, the greater efficiency of CC observed in Experiment 4 (when O1 was presented in both Phase 1 and Phase 2) cannot be explained in terms of more potent retroactive interference with the X-O1 association by the X-O2 association.

General Discussion

The goal of this series was to assess differences between extinction and CC in humans using a large number of participants to achieve sufficient statistical power. In summary, the following conclusions can be drawn: (a) the data indicate that the impact of a procedure on the expression of acquired memories likely differs depending on whether the memories assessed are more centrally emotional or cognitive. In the former case, CC has a stronger impact on performance than extinction; (b) in the latter case, CC is less effective than extinction at reducing outcome expectancy if the target outcome is presented at test, but (c) CC is more efficient if the target outcome is presented in both Phase 1 (target training) and Phase 2, perhaps because it attenuates reinstatement of Phase 1 when the target outcome is presented at test. This suggests greater sensitivity of CC than extinction to recovery effects, at least in the cognitive domain.

The impact of CC and extinction on outcome expectancy was massive, often corresponding to a Cohen's *d* larger than 2. Consistent with the literature, CC also impacted evaluative conditioning. If extinction also did, it did so to a lesser degree, a result consistent with the existing literature. But, even in the case of CC, the effect size on evaluative conditioning (Cohen's $d = 0.3$) appears to be considerably less than its effect on the outcome expectancy (see Figure 8). If we had used a more intense US, it is possible that the effect size for evaluative conditioning, and the potential for CC to attenuate evaluative conditioning, would have been larger; Experiment 3 indicated that the valence of the outcome had no impact on outcome expectancy. The IAPS pictures we used as USs effectively engage emotional circuits in the brain (i.e., Aldhafeeri, Mackenzie, Kay, Alghamdi, & Sluming, 2012) and consequently can be considered biologically relevant (a conclusion supported by the data in Figures 2 and 5), but they pale in comparison to the USs used in animal

studies (food for hungry animals, strong electric shocks). Mild non-invasive electric shocks are sometimes used in human studies but would not have been appropriate here because it would have been difficult to identify appetitive USs that match them in intensity. It is also possible that our dependent variable was not as sensitive as more physiological measures of emotional activation (such as the skin conductance response). In any case, it is unclear what effect size to expect for evaluative conditioning because, until recently, reporting effect size was not a common practice in psychological research. It would be surprising if less than an hour of occasional exposure to mildly emotional stimuli led to large changes in the likability of otherwise fairly meaningless stimuli. From that point of view, a small Cohen's *d* of 0.3 would be expected.

The observed effect sizes for the efficiency of CC over that of extinction were also quite small: Cohen's *d* for the differential impact of extinction and CC on evaluative conditioning was around 0.2, whereas Cohen's *d* for the differential impact of extinction and CC on the outcome expectancy was around 0.3. All told, one implication of the present study is that researchers in the future will need to pay more attention to statistical power and effect size when assessing the differential efficiency of CC and extinction. If the effect sizes observed here are representative, they pose serious limitations on the kind of conclusions that can be reached with low-powered studies. Paying more attention to effect size and the way they vary as a function of the dependent measure (for instance, valence rating on a Likert scale vs. emotional priming) might help to clarify those issues.

A key conclusion from our study is the dissociation between different measures of conditioning. Although all measures presumably are dependent on memory, we viewed measures of outcome expectancy as reflecting a more cognitive assessment of the cue-outcome relation, and measures of evaluative conditioning as being more dependent on emotional processing of the cue. This is hardly an innovation. Konorski (1967) long ago suggested that USs have both sensory and affective properties and that a CS might associate with these different attributes independently. This view was explored further in a later version of Wagner's SOP model (AESOP: Wagner & Brandon, 1989) and has been exploited in empirical works (i.e. Bakal, Johnson, & Rescorla, 1974; Betts,

Brandon & Wagner, 1996: Delamater, 2012; Ganesan & Pearce, 1988). It is echoed, but with a somewhat different focus, in the dual-process account of conditioning (i.e., McLaren et al., 2014), according to which the outcome of conditioning in humans can best be understood as relying on two processes, one best explained by associative models, the other by propositional ones (see also the system I/system II dichotomy in behavioral economy, Kahneman, 2011). On a continuum going from hot emotions to cold cognition, one could view the associative process as being closer to the hot emotional pole, while the propositional process would be closer to the cold cognitive one. One could also argue that, rather than reflecting two processes, patterns such as the one reported here reflect a single source of knowledge the effect of which on behavior depends on how it is probed by the question asked of the participant (i.e., Vadillo, Miller, & Matute, 2005; Whittlesea & Price, 2001). This is a complex issue which, we think is not wholly empirical because it hinges on one's definition of an association. Given some definitions, saying that there is a single memory that can affect behavior differentially depending on how it is queried, and saying that independent associations between the CS and various properties of the US are created is actually saying the same thing, particularly if we assume that a question will preferentially activate some of the CS-US associations over others: The latter formulation is potentially a way of formalizing the former one. In any case, no matter how the dissociation between emotional and cognitive measures is conceptualized, the bottom line remains the same: When talking of the relative efficiency of one treatment (i.e., CC) over another (i.e., extinction), we must pay attention to the dependent variables used to evaluate the efficiency of each treatment because the results might depend on which processes are tapped by that dependent variable.

What are the implications of the present results for our understanding of the mechanisms underlying counterconditioning and extinction? We can only speculate. From an associative interference point of view, one might conclude: (a) the Phase 1 treatment establishes a cue-outcome association (A1); (b) The Phase 2 treatment establishes an alternative cue-outcome association (A2); (c) A2 will interfere with the expression of A1 based on the similarity between the outcomes involved in A1 and A2. The more similar they are, the less efficient the interference. Moreover,

following Konorski (1967), a cue enters in association separately for the affective and the sensory properties of an outcome. Evaluative conditioning primarily reflects the association with the affective properties, whereas the outcome expectancy reflects the association with the sensory properties. Moreover, in the case of extinction, the Phase 2 outcome is the absence of the Phase 1 outcome (see Pearce & Hall, 1980, for a further discussion on this complex topic).

Those principles potentially explain most of the results observed in the present study. Arguably, the sensory properties of the Phase 2 outcome in CC are more similar to the sensory properties of the Phase 1 outcome than in extinction because O2 is more like O1 than is merely the lack of O1. This would explain the greater efficiency of extinction than CC with regard to outcome expectancy. The reverse would be true regarding the affective properties of Phase 1 and Phase 2 outcomes, hence the greater efficiency of CC than extinction when it comes to evaluative conditioning.

But what is to be made of the results of Experiment 4, where the introduction of O1 in Phase 2 made CC more efficient than extinction at impacting outcome expectancy? When we designed that experiment and as stated in the rationale for it, we hypothesized that the mention of O1 in the test question might lead to some reinstatement and that, for unknown reasons, CC was more susceptible to reinstatement than was extinction. The analysis above provides an alternative explanation. The presentation of O1 and O2 during Phase 2 would draw attention to their distinctive sensory attributes, thereby enhancing the potential of the association between X and the sensory properties of O2 to interfere with the expression of the association between X and the sensory properties of O1.

References

Aldhafeeri, F. M., Mackenzie, I., Kay, T., Alghamdi, J., & Sluming, V. (2012). Regional brain responses to pleasant and unpleasant IAPS pictures: Different networks. *Neuroscience Letters, 512*, 94-98.

Algina, J., & Kesselman, H. J. (2003). Approximate confidence intervals for effect size. *Educational and Psychological Measurements, 63,* 537-553.

Bakal, C. W. Johnson, R. D., & Rescorla, R. A. (1974). The effect of a change in US quality on the blocking effect. *Pavlovian Journal of Biological Science, 9,* 97-103.

Betts, S. L. Brandon, S.E., & Wagner, A. R. (1996). Differential blocking of the acquisition of conditioned eyeblink responding and conditioned fear with a shift in US locus. *Animal Learning & Behavior*, *24*, 459-470.

Bouton, M. E. (2017). Extinction: Behavioral mechanisms and their implications. In Byrne, G. (Ed), *Learning and memory: A comprehensive reference, 2nd edition: Vol 1: Learning theory and behavior* (pp. 61-83). Cambridge, MA: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Routledge.

Crump, M. J. C., Hannah, S. D., Allan, L. G., & Hord, L. K. (2007). Contingency judgments on the fly. *Quarterly Journal of Experimental Psychology, 60,* 753-761.

Cummings, G. (2012). *Understanding the new statistics: effect size, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Delamater, A. R. (2012). On the nature of CS and US representation in Pavlovian learning. *Learning & Behavior, 40,* 1-23.

De Jong, P. J., Vorage, I., & Van den Hout, M. A. (2000). Counterconditioning in the treatment of spider phobia: Effects on disgust, fear, and valence. *Behaviour Research and Therapy, 38*, 1055-1069.

Dunsmoor, J. E., Campese, V. D., Ceceli, A. O., LeDoux, J. E., & Phelps, E. A. (2015). Novelty-facilitated extinction: providing a novel outcome in place of an expected threat diminishes recovery of defensive responses. *Biological Psychiatry, 78,* 203-209.

Dunsmoor, J. E., Kroes, M. C. W., Li, J., Daw, N. D., Simpson, H. B., & Phelps, E. A. (2019). Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. *Journal of Neuroscience, 39,* 3264-3276.

Engelhard, I. M., Leer, A., Lange, E. & Olatuju, B. O. (2014). Shaking the icky feeling: Effects of extinction and counterconditioning on disgust-related evaluative learning. *Behavior therapy, 45*, 708-719.

Escobar, M., Arcediano, F., & Miller, R. R. (2001). Conditions favoring retroactive interference between antecedent events (cue competition) and between subsequent events (outcome competition). *Psychonomic Bulletin & Review, 8*, 691-697.

Ganesan, R., & Pearce, J. M. (1988). Effects of changing the unconditioned stimulus on appetitive blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 280-291.

Hannah, S. D., Crump, M. J. C., Allan, L. G., & Siegel, S. (2009). Cue-interaction effects in contingency judgments using the streamed-trial procedure. *Canadian Journal of Experimental Psychology, 63,* 103-112.

Holmes, N. M., Leung, H. T., & Westbrook, R. F. (2016). Counterconditioned fear responses exhibit greater renewal than extinguished fear responses. *Learning and Memory, 23*, 141-150.

Kang, S., Vervliet, B., Engelhard, I. M., van Dis, E. A. M., & Hagenaars, M. A. (2018). Reduced return of threat expectancy after counterconditioning versus extinction. *Behavior Research and* Therapy, 108, 78-84.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus, and Giroux.

Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning: an effective technique for changing conditioned preferences. *Experimental Psychology, 58*, 31-38.

Konorski, J. (1967). *Integrative activity of the brain*. Chicago: Chicago University Press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system

(IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

Laux, J. P., Goedert, K. M., & Markman, A. B. (2010). Causal discounting in the presence of a stronger cue is due to bias. *Psychonomic Bulletin & Review, 17,* 213-218.

Lucas, J., Luck, C. C., & Lipp, O. V. (2018). Novelty-facilitated extinction and the reinstatement of conditional human fear. *Behaviour Research and Therapy, 109,* 68-74.

Maia, S., Lefèvre, F., & Jozefowiez, J. (2018). Psychophysics of associative learning: Quantitative properties of subjective contingency. *Journal of Experimental Psychology: Animal Learning & Cognition, 44,* 67-81.

McLaren, I. P. L, Forrest, R. P., McLaren, F. W., Jones, M. R. F., Aitken, N. J., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of Learning and Memory, 108,* 185-195.

Meulders, A., Karsdop, P. A., Claes, N. & Vlaeyen, J. W. (2015). Comparing counterconditioning and extinction as methods to reduce fear of movement-related pain. *Journal of Pain, 16*, 1353-1365.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

Polack, C. W., Jozefowiez, J., & Miller, R. R. (2017). Stepping back from "persistence and relapse" to see the forest: Associative interference. *Behavioural Processes*, *141*, 128-136.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned not of unconditioned stimuli. *Psychological Review, 87,* 532-552.

Peirce, J. W. (2007). Psychopy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8-13.

Raes, A. K., & De Raed, R. (2012). The effect of counterconditioning on evaluative responses and harm expectancy in a fear conditioning paradigm. *Behavior Therapy, 43,* 757-767.

Reynolds, G., Field, A. P., & Askew, C. (2018). Reductions in children's vicariously learnt avoidance and heart responses using positive modeling. *Journal of Clinical Child and Adolescent Psychology, 47*, 555-558.

Siegel, S., Allan, L. G., & Hannah, S. D. (2009). Applying signal detection theory to contingency assessment. *Comparative Cognition & Behavior Reviews, 4,* 116-134.

Tunstall, B. J., Verendeev, A., & Kearns, D. N. (2012). A comparison of therapies for the treatment of drug cues: Counterconditioning vs. extinction in male rats. *Experimental and Clinical Psychopharmacology, 20,* 447-453.

Vadillo, M. A., Miller, R. R., & Matute, H. (2005). Causal and predictive-value judgments, but not predictions, are based on cue-outcome contingency. *Learning & Behavior, 33,* 172-183.

VanElzakker, M. B., Dahlgren, M. K., Davis, F. C., Dubois, S., & Shin, L. M. (2014). From Pavlov to PTSD: The extinction of conditioned fear in rodents, humans, and anxiety disorders. *Neurobiology of Learning and Memory, 113,* 3-18.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology, 9,* 215-248.

Wagner, A. R., & Brandon, S. E. (1989). Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). In S. B. Klein & R. R. Mowrer (Eds), Pavlovian conditioning and the status of traditional learning theory (pp. 149-189). Hillsdale, NJ: Erlbaum.

Whittlesea, B. W. A., & Price, J. R. (2001). Implicit/explicit memory versus analytic/nonanalytic processing: Rethinking the mere exposure effect. *Memory & Cognition, 29,* 234-246.

Appendix: Gender differences in associative learning in the cognitive and emotional domain

This appendix documents Cohen's *d* for gender differences. We used the formula recommended by Cummings (2012) for independent-group design

$$d = \left[1 - \frac{3}{4(n_F + n_M - 2) - 1}\right] \frac{m_F - m_M}{\sqrt{\frac{(n_F-1)s_F^2+(n_M-1)s_M^2}{n_F+n_M-2}}}$$

where n_F (respectively n_M) is the number of female (respectively male) participants; m_F (respectively m_M) is the group mean for the female (respectively male) participants; s_F^2 (respectively s_M^2) is the variance for the female (respectively male) participants. 95% confidence intervals for Cohen's *d* were computed using ESCI (https://thenewstatistics.com/itns/esci/), following the method described in Cummings (2012).

Figure A1 shows the effect of gender on the outcome valence rating in Experiments 1 and 2 (Figures 2 and 5 in the core of the paper). It seems that, overall, female participants tended to rate the positive outcome higher and the negative one lower than their male counterpart.

Figure A2 shows the effect of gender on the valence rating of the target cue in Experiments 1 and 2 (Figures 3 and 6 in the core of the paper). There are no clear differences, which might reflect the overall lack of sensitivity of that dependent variable.

The top panel of Figure A3 shows the effect of gender on the outcome expectancy rating in Experiments 1 to 4 (top panel of Figures 4, 7, 9, and 11). There seems to be an overall tendency for female participants to rate the X-O1 relation higher than the male participants. The effect is clearer in Experiment 1 but cannot be ruled out for the other studies. Figure A4 shows the corresponding data for Experiment 5 (corresponding to the top panel of Figure 13 in the core of the text. Data from Experiment 5 are presented apart from the other ones as it involved assessment of the X-O2 relation instead of the X-O1 relation), It reveals no detectable gender difference.

The bottom panel of Figure A4 shows the effect of gender on the efficiency of extinction and counterconditioning in Experiments 1 to 4 (bottom panel of Figures 4, 7, 9, and 11). Extinction and counterconditioning are less effective at impacting outcome expectancy rating in female participants

relative to male participants in the two experiments that used emotionally charged pictures as outcome (Experiments 1 and 2). No similar effect can be detected in Experiments 3 and 4, nor in the equivalent data for Experiment 5 (displayed in the bottom panel of Figure A4, corresponding to the bottom panel of Figure 13), which all used neutral pictures as outcomes.

Author Notes

Jérémie Jozefowiez, Univ.Lille, CNRS, UMR 9193 – SCALab – Sciences Cognitives et Affectives, F-59000 Lille France; Cody W. Polack, Alaina S. Berruti, Yaroslav Moshchenko, Tori Peña, and Ralph R. Miller, Department of Psychology, SUNY-Binghamton (Binghamton, New York, USA).

We would like to thank Yoan Villemin for his help in collecting and analyzing the data.

Correspondence concerning this article should be addressed to Jérémie Jozefowiez, Laboratory of Affective and Cognitive Sciences (SCALab UMR CNRS 9193), Université de Lille, Campus de Lille SHS, Domaine Universitaire du Pont de Bois, BP 60149, 58653 Villeneuve d'Ascq Cedex, France. E-mail: jeremie.jozefowiez@univ-lille.fr. This research was supported in part by NIH Award MH033881. All raw data and computer code for the experiment is available upon request from either J. Jozefowiez or R. R. Miller. The authors report no conflicts of interest.

Table 1

Identify, arousal score and valence score of the IAPS pictures used as outcomes in Experiments 1 and 2.

Category	IAPS identifier	Arousal Score	Valence Score
Human beings	Smiling children #2347	mean: 5.56, sd: 2.34	mean: 7.83, sd: 1.36
Human beings	Bloodied face #3051	mean: 5.62, sd: 2.45	mean: 2.30, sd: 1.86
Animals	Baby cheetahs #1722	mean: 5.22, sd: 2.49	mean: 7.04, sd: 2.02
Animals	Dead dog #9185	mean: 5.65, sd: 2.35	mean: 1.97, sd: 1.16
Objects	Ice cream #7330	mean: 5.14, sd: 2.58	mean: 7.69, sd: 1.84
Objects	Dirty toilet #9301	mean: 5.28, sd: 2.46	mean: 2.26, sd: 1.56

Table 2

Composition of the trial streams during warmup in Experiments 1, 2, and 3. 'n A-B' means that n trials of stimuli A and B were presented during each cycle.' n A-' means that n trials of A alone was presented during each cycle. The program cycled once through Phase 1 and twice through Phase 2 before the outcome expectancy question was presented. During each cycle, the order of presentation of the trials was determined randomly. X and O1 were the target cues and outcomes.

Condition	Phase 1 (one cycle)	Phase 2 (two cycles)
100% training	6 X-O1 / 5 Y-	6 Y- / 5 Y-O2
50% training	3 X-O1 / 3 X- / 3 Y-O2 / 2 Y-	3 X-O1 / 3 X- / 3 Y-O2 / 2 Y-
0% training	6 X- / 5 Y-O2	6 Y- / 5 Y-O2

Table 3

Composition of the trial streams during training in Experiments 1, 2, and 3. n A-B means that n trials of stimuli A and B were presented together during each cycle. n A- means that n trials of A alone were presented during each cycle. The program cycled once through Phase 1 and twice through Phase 2 before the outcome expectancy question was presented. During each cycle, the order of presentation of the trials was determined randomly. X and O1 are the target cues and outcomes.

Condition	Phase 1 (one cycle)	Phase 2 (two cycles)
Control	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 Y-O2
Extinction	5 X-O1 / 1 X- / 5 Y-	1 Y- / 5 X- / 5 Y-O2
Counterconditioning	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 X-O2

Table 4

Composition of the trial streams during the warmup of Experiments 4 and 5. n A-B means that n trials of stimuli A and B were presented together during each cycle. n A- means that n trials of A alone were presented during each cycle. The program cycled once through Phase 1 and twice through Phase 2 before the outcome expectancy question was presented. During each cycle, the order of presentation of the trials was determined randomly. X and O1 are the target cues and outcomes.

Condition	Phase 1 (one cycle)	Phase 2 (two cycles)
100% training	6 X-O1 / 5 Y-	6 Y- / 6 Y-O2 / 5 O1
50% training	3 X-O1 / 3 X- / 3 Y-O2 / 2 Y-	3 X-O1 / 3 X- / 3 Y-O2 / 2 Y- / 5 O1
0% training	6 X- / 5 Y-O2	6 Y- / 6 Y-O2 / 5 O1

Table 5

Composition of the trial streams during training in Experiments 4 and 5. n A-B means that n trials of stimuli A and B were presented together during each cycle. n A- means that n trials of A alone were presented during each cycle. The program cycled once through Phase 1 and twice through Phase 2 before the outcome expectancy question was presented. During each cycle, the order of presentation of the trials was determined randomly. X and O1 are the target cues and outcomes.

Condition	Phase 1 (one cycle)	Phase 2 (two cycles)
Control	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 Y-O2
Extinction	5 X-O1 / 1 X- / 5 Y-	1 Y- / 5 X- / 5 Y-O2
Counterconditioning	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 X-O2
Control – O1	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 Y-O2 / 5 O1
Extinction – O1	5 X-O1 / 1 X- / 5 Y-	1 Y- / 5 X- / 5 Y-O2 / 5 O1
Counterconditioning – O1	5 X-O1 / 1 X- / 5 Y-	6 Y- / 5 X-O2 / 5 O1

Figure 1. Top panel: Layout showing how the cues (X and Y) and the outcomes (O1 and O2) were presented during training. Bottom panel: Example of the outcome question asked to the participants during testing.

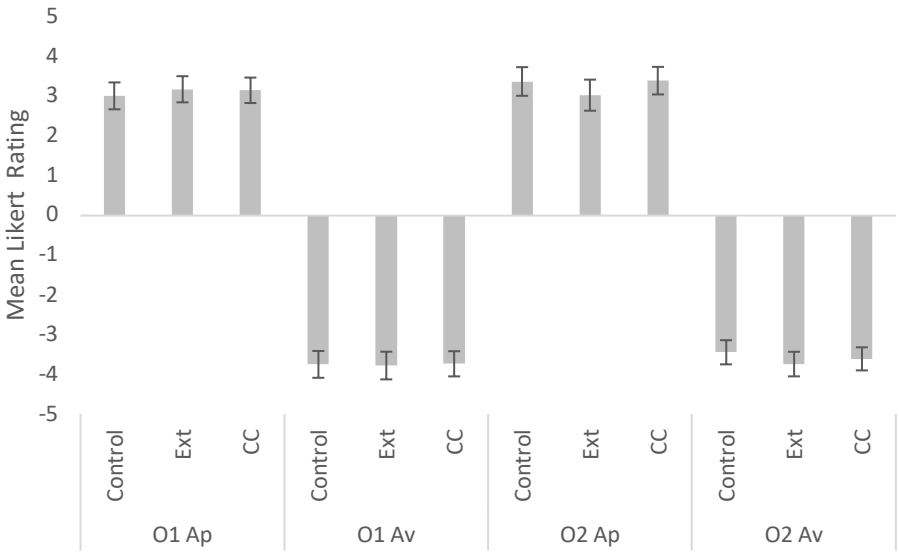

Figure 2. Mean Likert rating for the emotional valence of the O1 and O2 as a function of condition in Experiment 1. Error-bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

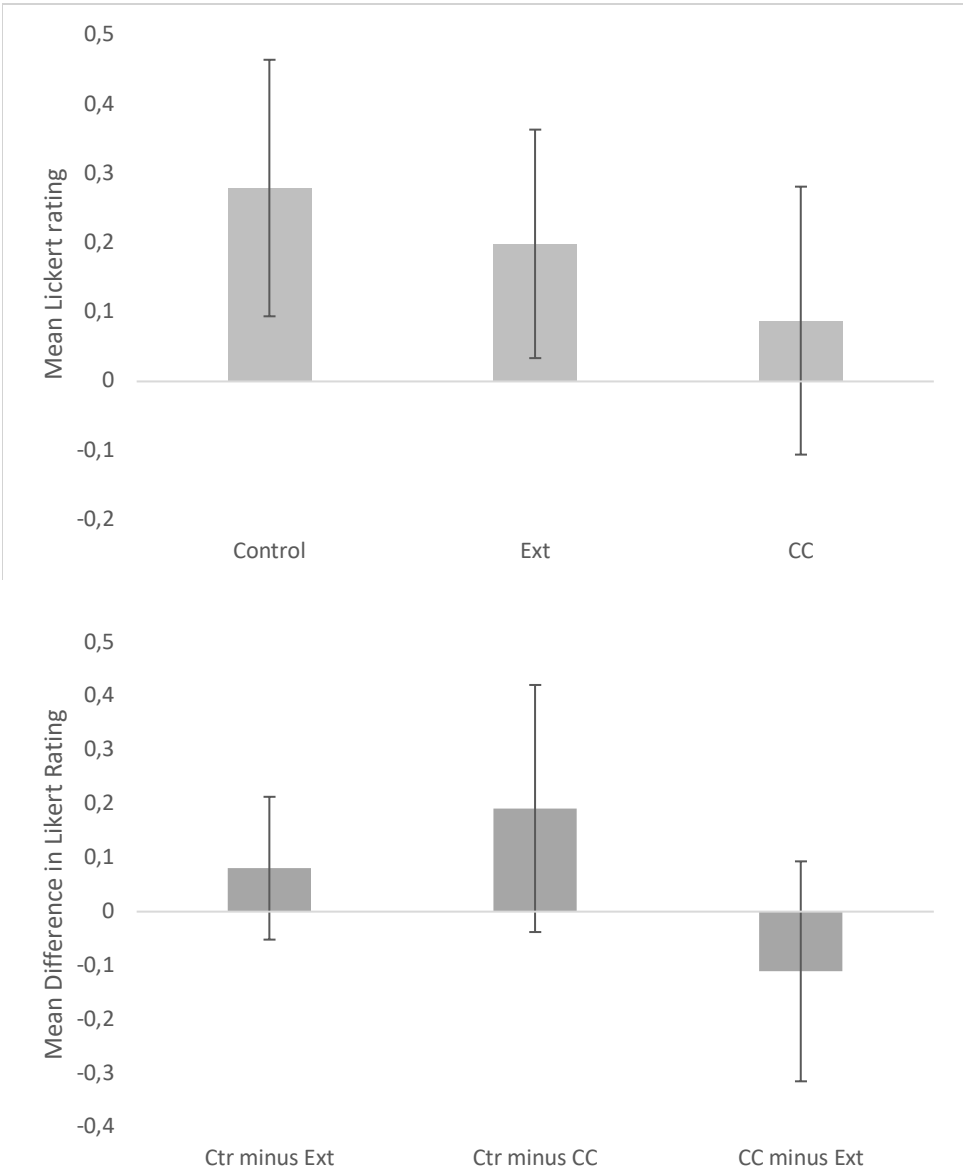

Figure 3. Top panel: Mean Likert rating for the emotional valence of cue X as a function of condition in Experiment 1. Bottom panel: Mean difference in Likert rating for the emotional valence of cue X for each pair of conditions in Experiment 1. Error-bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

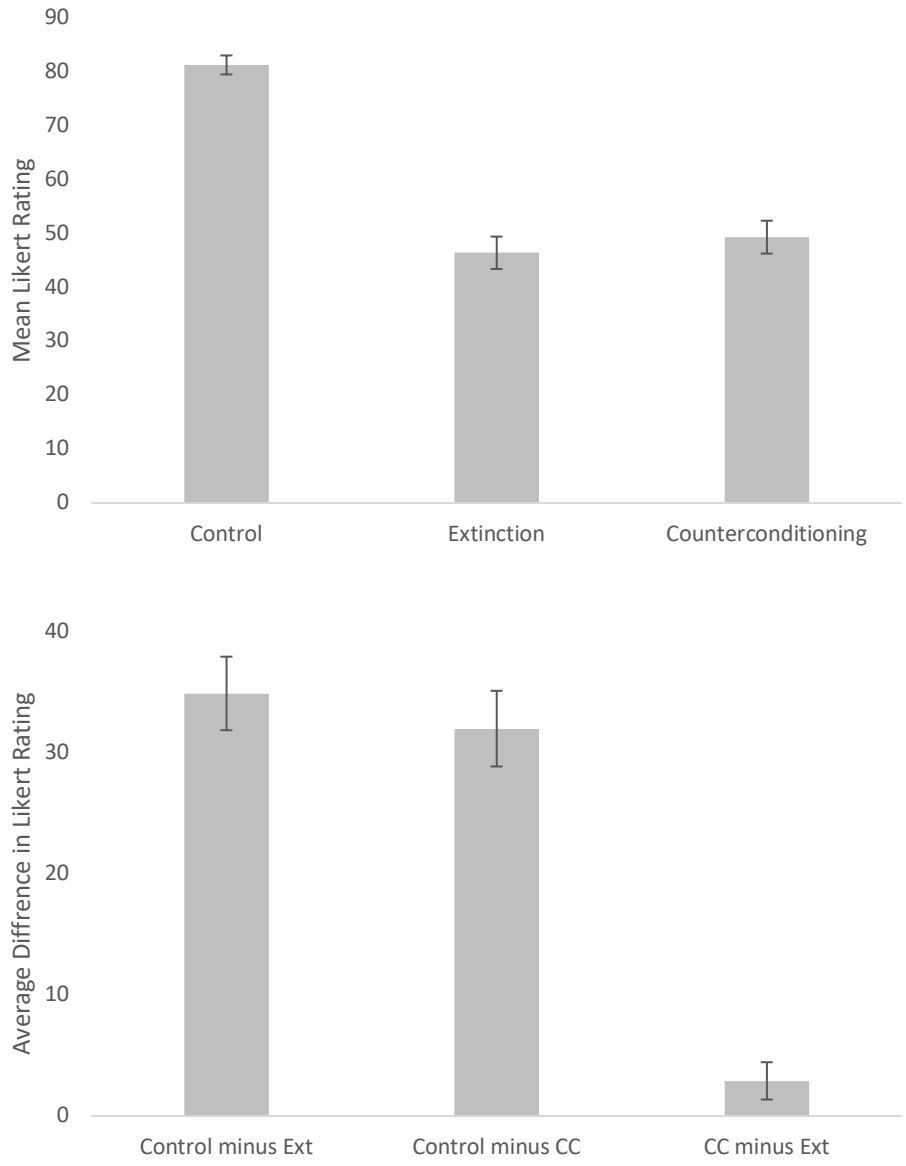

Figure 4. Top panel: Mean Likert rating for the O1 expectancy in the presence of X as a function of condition in Experiment 1. Bottom panel: Mean difference in the Likert rating for the O1 expectancy in the presence of X for each pair of conditions. Error bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

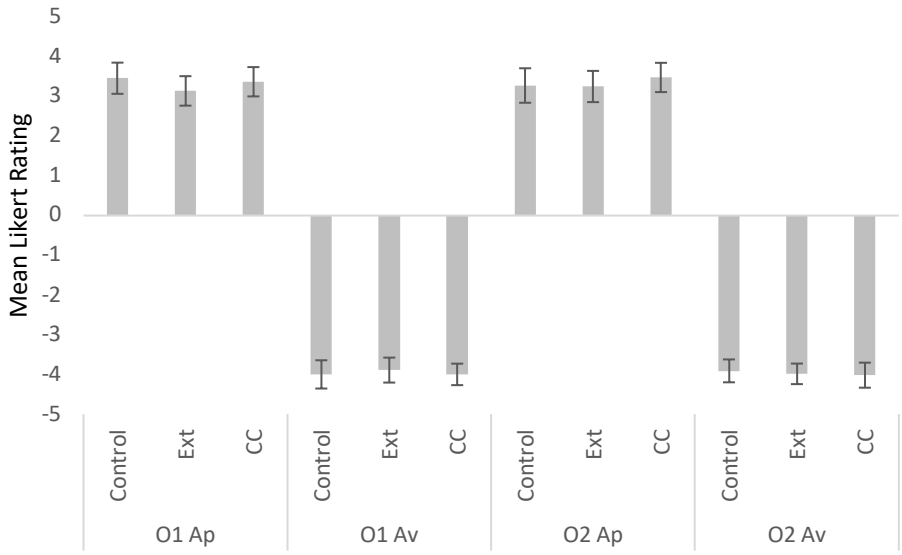

Figure 5. Mean Likert rating for the emotional valence of O1 and O2 in Experiment 2 as a function of condition in Experiment 2. Error-bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

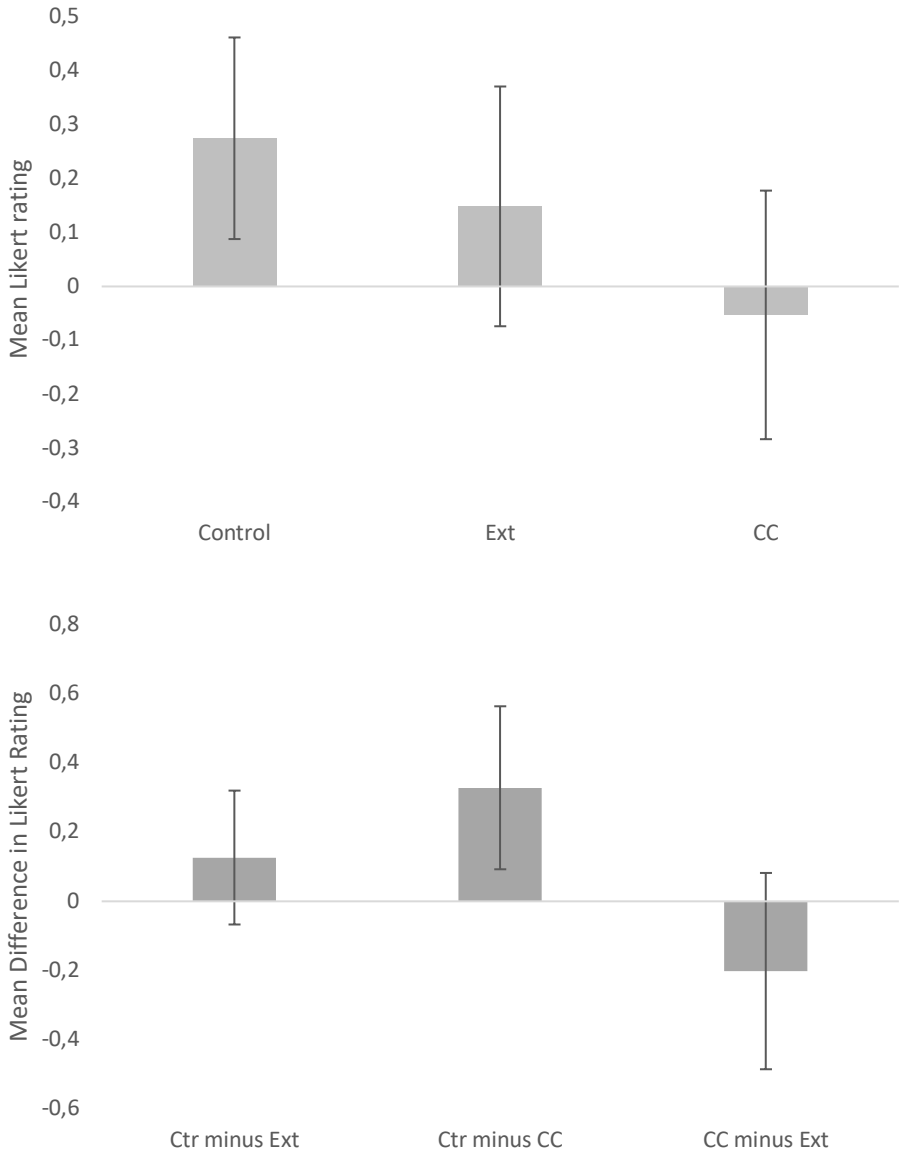

Figure 6. Top panel: Mean Likert rating for the emotional valence of cue X as a function of condition in Experiment 2. Bottom panel: Mean Likert rating for the emotional valence of cue X for each pair of conditions in Experiment 1. Error-bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

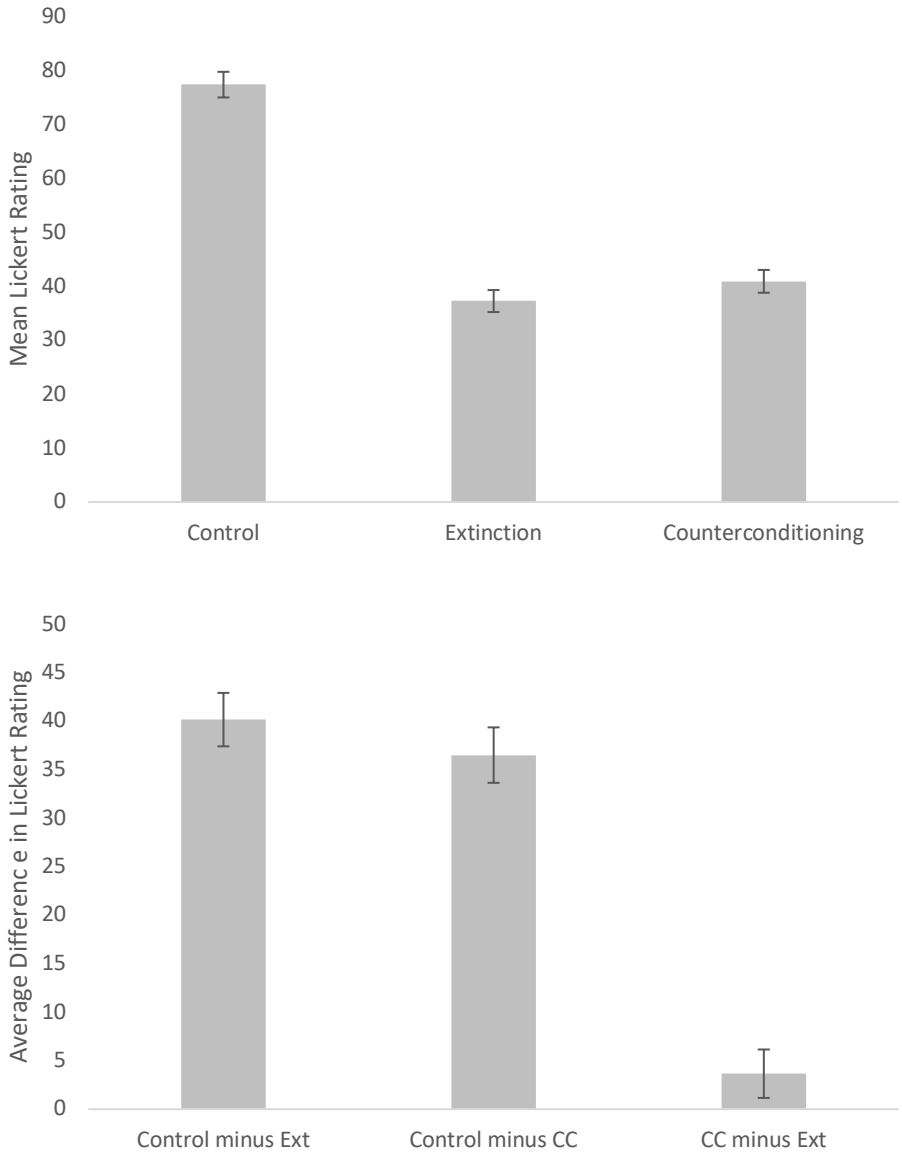

Figure 7. Top panel: Mean Likert rating for the O1 expectancy in the presence of X as a function of condition in Experiment 2. Bottom panel: Mean difference in the Likert rating for the O1 expectancy in the presence of X for each pair of conditions. Error bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

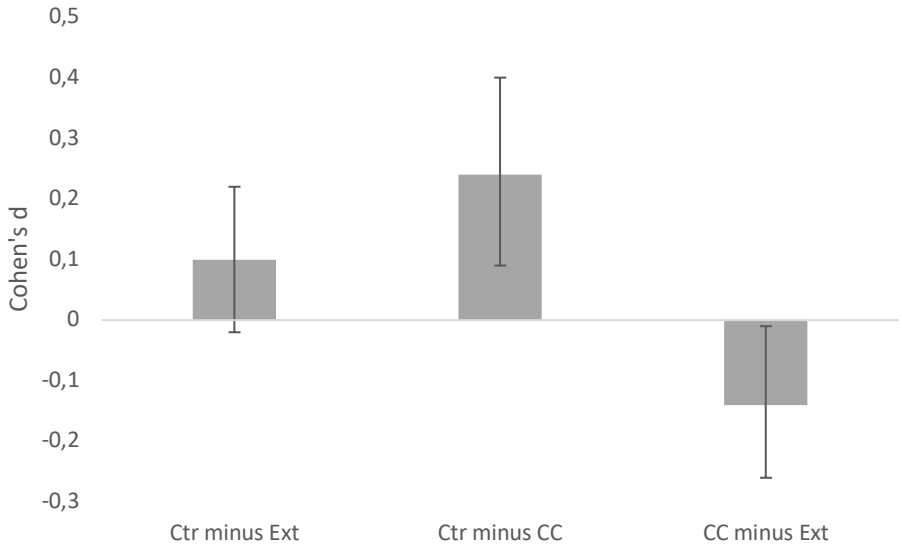

Figure 8. Effect size for evaluative conditioning across Experiments 1 and 2. Error-bars are 95% CI.

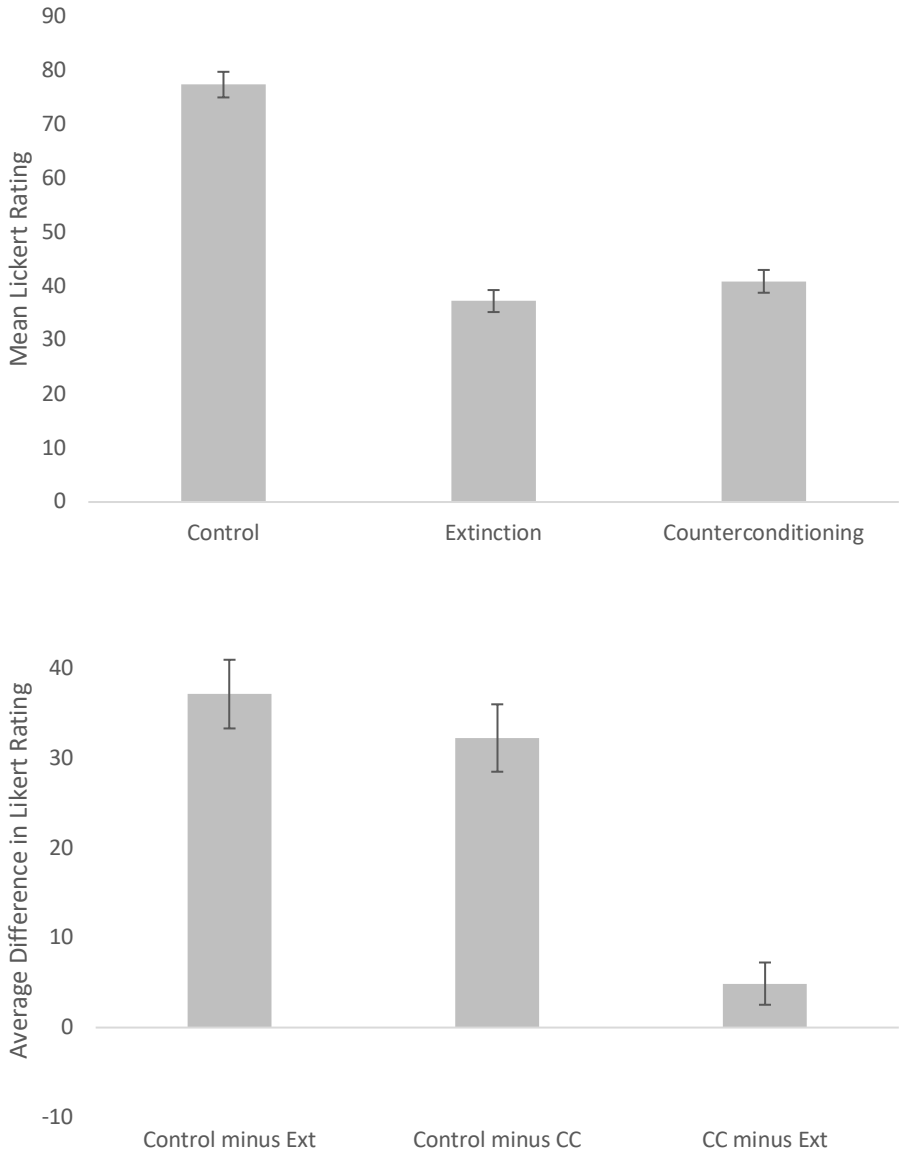

Figure 9. Top panel: Mean Likert rating for the O1 expectancy in the presence of X as a function of condition in Experiment 3. Bottom panel: Mean difference in the Likert rating for the O1 expectancy in the presence of X for each pair of conditions. Error bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

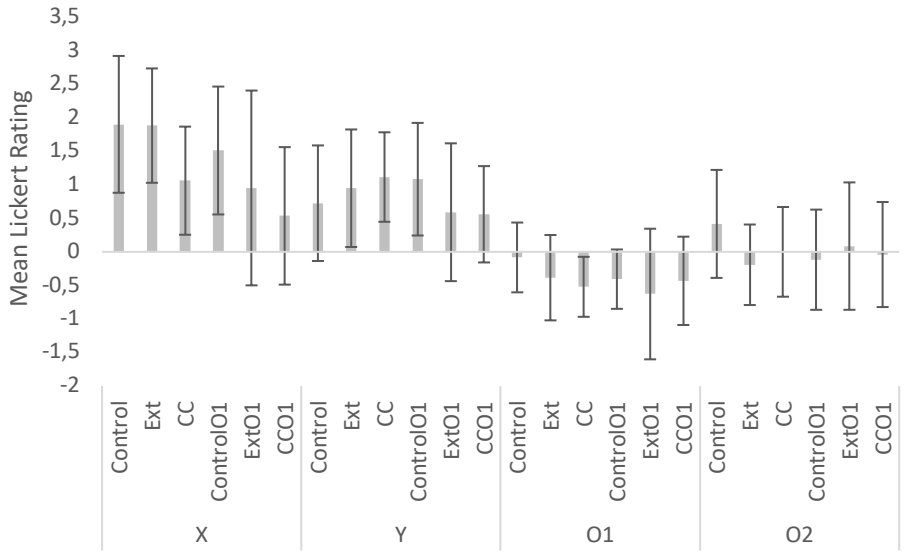

Figure 10. Mean Likert ratings for the emotional valence of the cues and outcomes in Experiment 4 as a function of condition. Error-bars are 95% CIs.

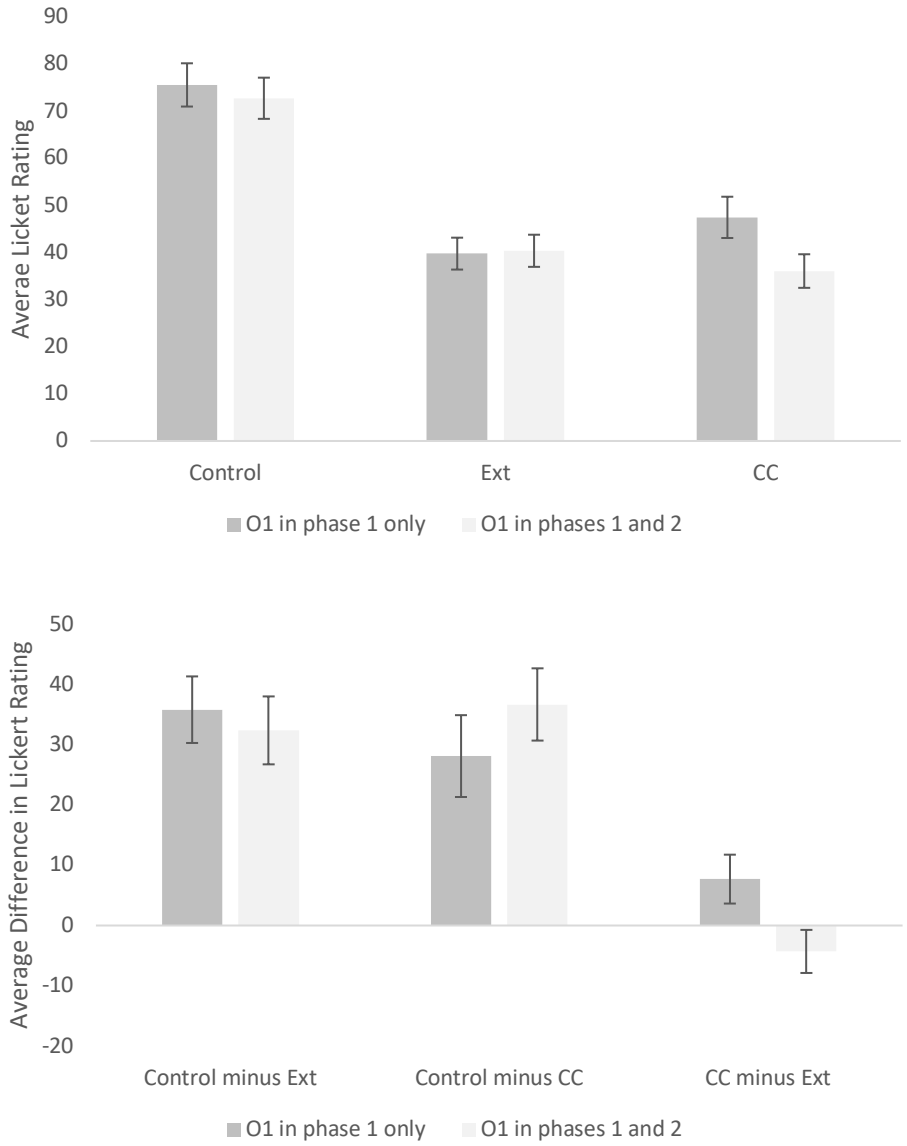

Figure 11. Top panel: Mean Likert ratings for the O1 expectancy in the presence of X as a function of condition in Experiment 4. Bottom panel: Mean difference in the Likert rating for the O1 expectancy in the presence of X for each pair of conditions. Error bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

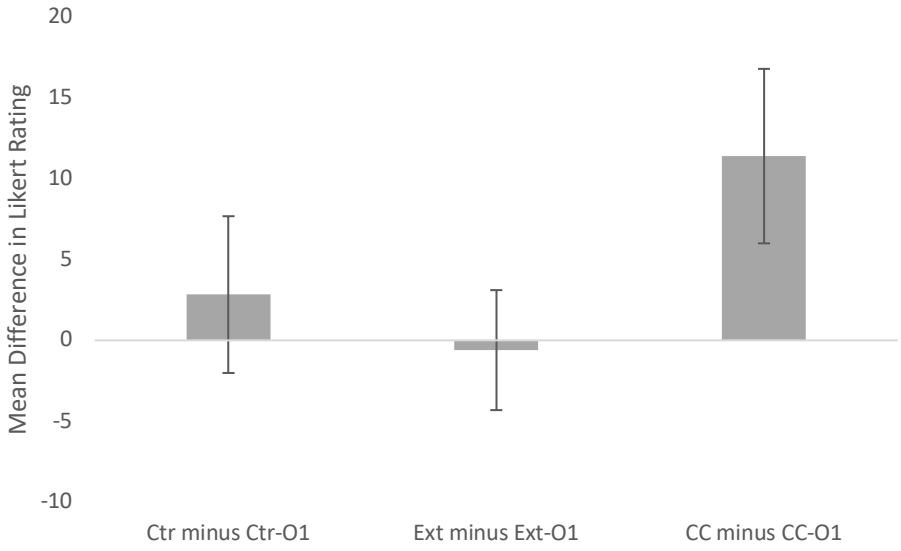

Figure 12. Mean difference in the Likert ratings for the O1 expectancy in the presence of X for each pair of conditions differing only in whether O1 was presented or not in Phase 2 in Experiment 4. Error bars are 95% CIs. Ctr = Control, Ext = Extinction, CC = counterconditioning.

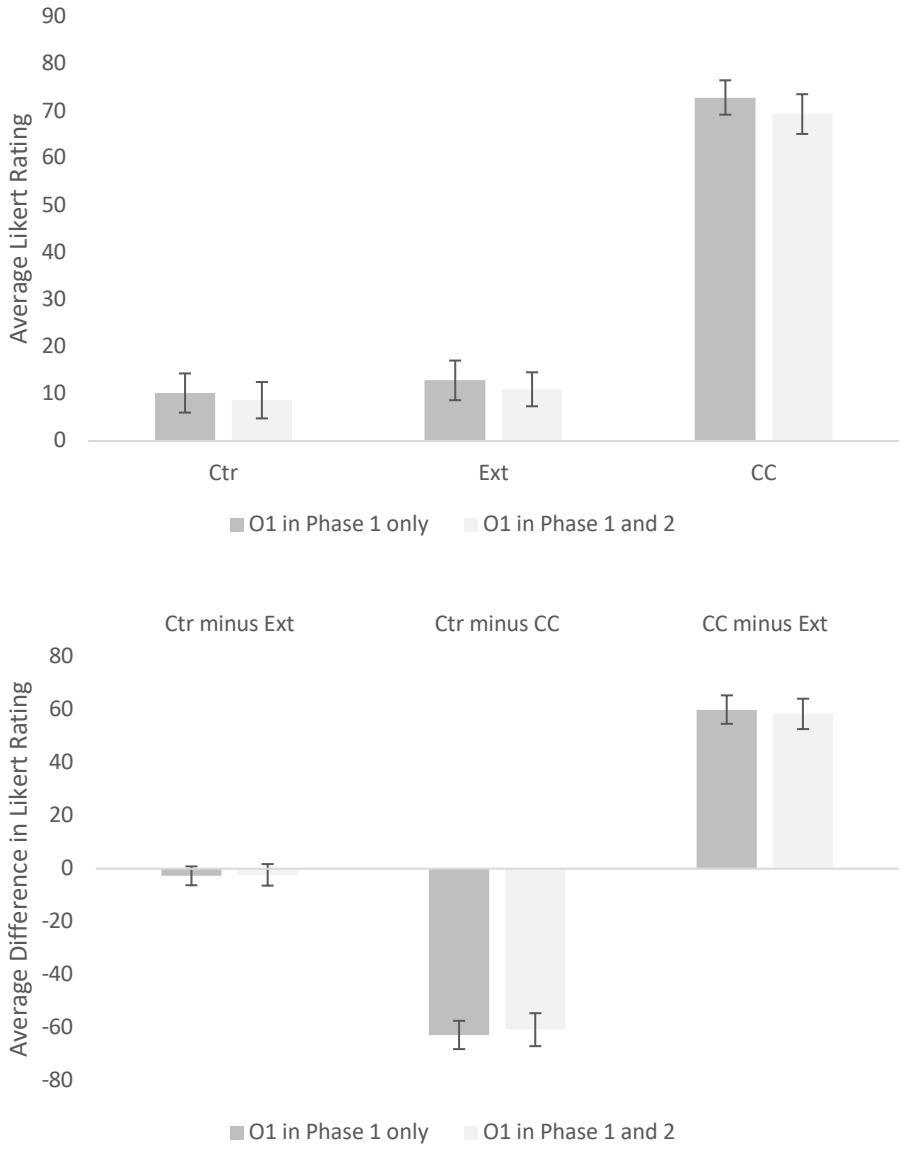

Figure 13. Top panel: Mean Likert ratings for the O2 expectancy in the presence of X as a function of condition in Experiment 5. Bottom panel: Mean difference in the Likert rating for the O2 expectancy in the presence of X for each pair of conditions. Error bars are 95% CIs. Ext = Extinction, CC = counterconditioning.

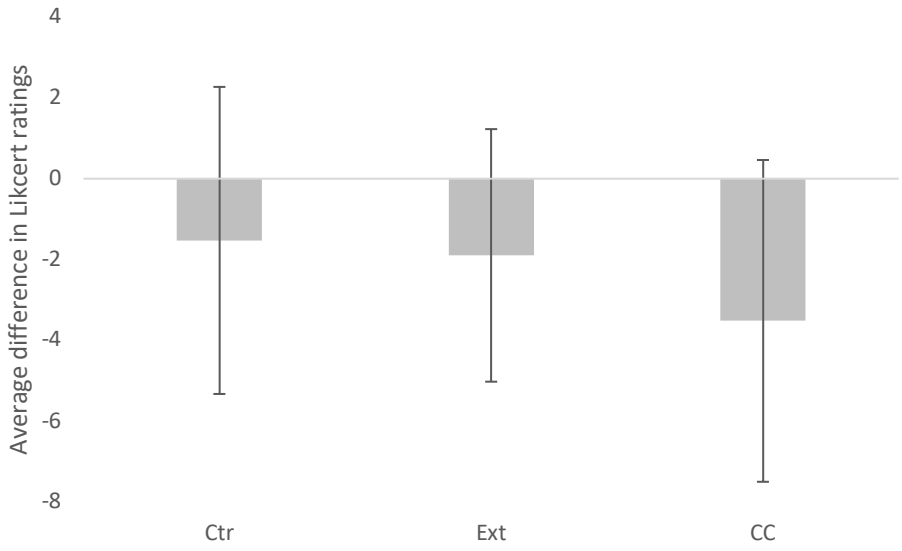

Figure 14. Mean difference in the Likert ratings for the the O2 expectancy in the presence of X for each pair of conditions differing only in whether O1 was presented or not in Phase 2 in Experiment 5. Error bars are 95% CIs. Ctr = Control, Ext = Extinction, CC = counterconditioning.

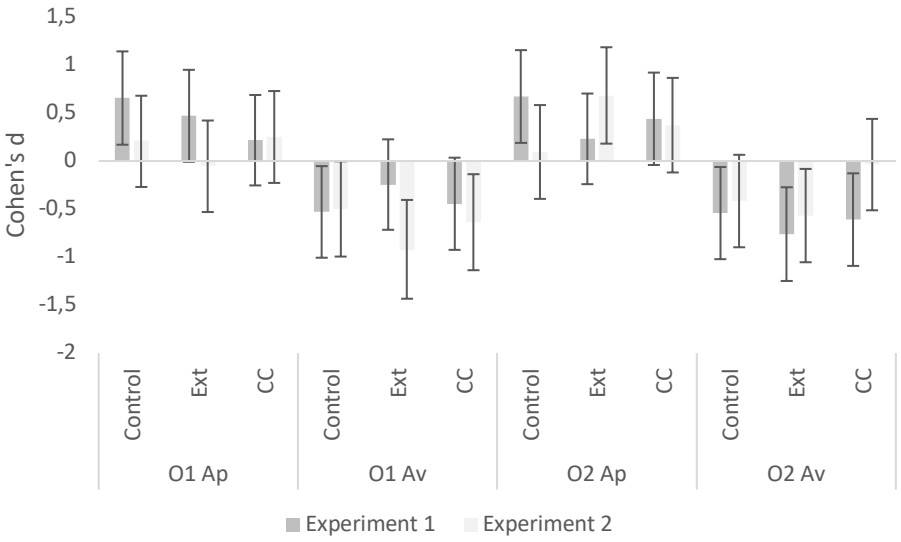

Figure A1. Cohen's d for the effect on gender on the outcome valence rating in Experiments 1 and 2. Error-bars are 95% CI.

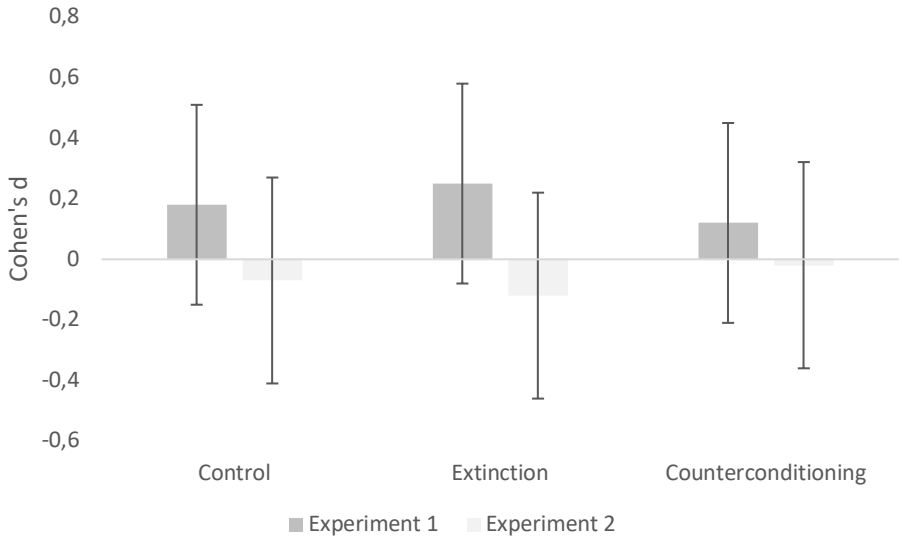

Figure A2. Cohen's d for the effect of gender on the cue valence rating as a function of condition in Experiments 1 and 2. Error-bars are 95% CI.

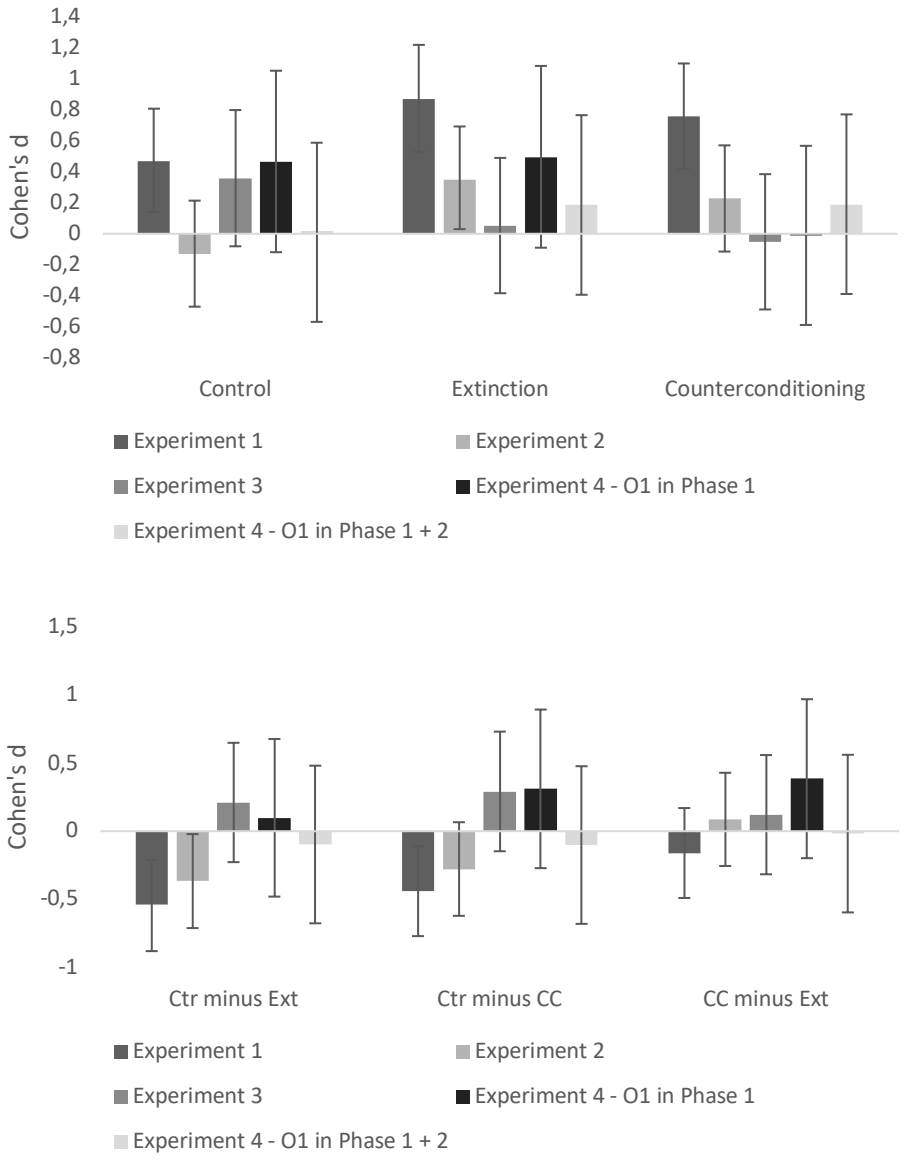

Figure A3. Top panel: Effect of gender on the outcome expectancy rating as a function of conditions in Experiments 1 to 4. Bottom panel: Effect of gender of the efficiency of extinction and counterconditioning at impacting outcome expectancy judgement in Experiments 1 to 4. Error-bars are 95% CI.

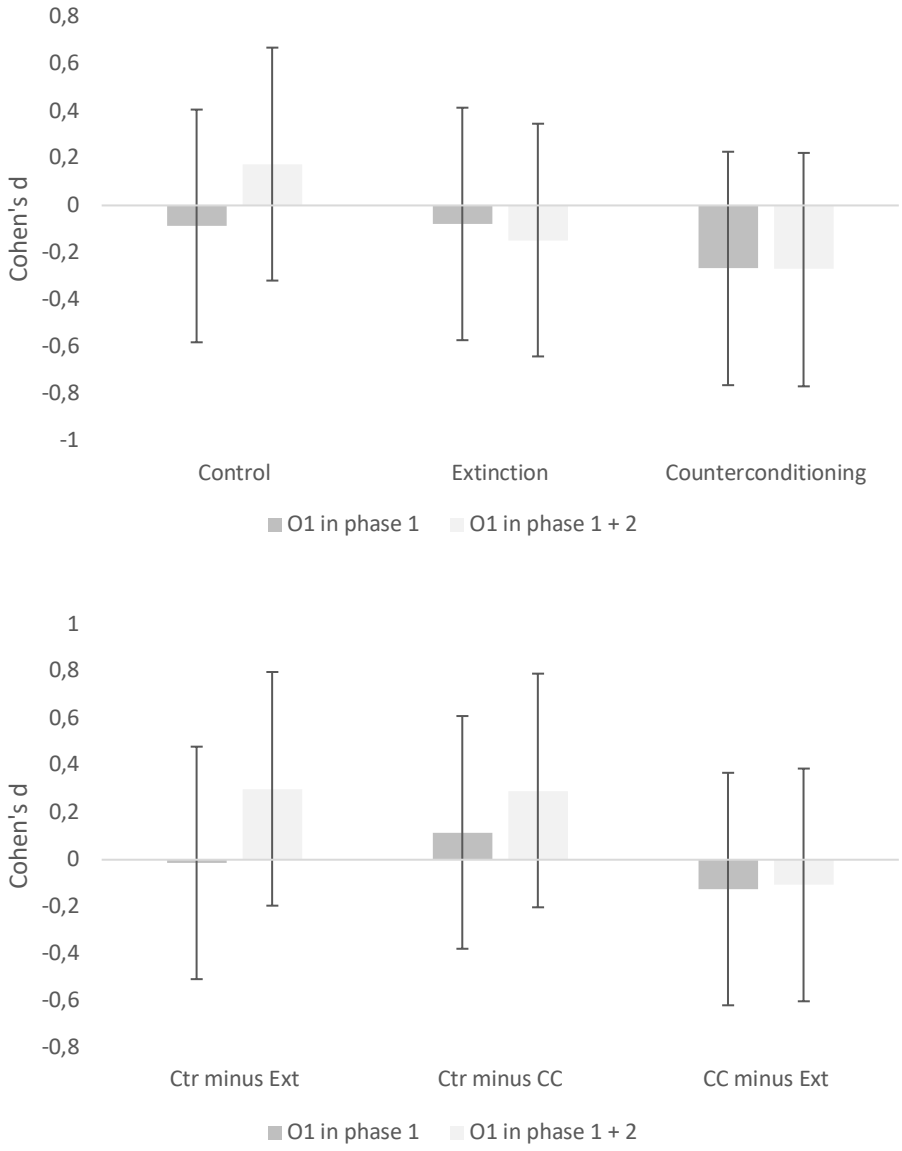

Figure A4. Top panel: Effect of gender on the outcome expectancy rating as a function of conditions in Experiment 5. Bottom panel: Effect of gender of the efficiency of extinction and counterconditioning at impacting outcome expectancy judgement in Experiment 5. Error-bars are 95% CI.