RUNNING HEAD: EVALUATIVE CONDITIONING

The Influence of Extinction and Counterconditioning Instructions on Evaluative Conditioning Effects

Anne Gast and Jan De Houwer

Ghent University

In press at: Learning and Motivation

Author note

Anne Gast and Jan De Houwer, Department of Experimental Clinical and Health Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium. E-mail addresses: anne.gast@uni-koeln.de, Jan.DeHouwer@UGent.be. Correspondence concerning this article should be addressed to Anne Gast.

The preparation of this paper was made possible by Methusalem Grant BOF09/01M00209 of Ghent University awarded to Jan De Houwer.

Abstract

In three experiments, we tested the influence of instructions about an allegedly upcoming extinction or counterconditioning phase on evaluative conditioning (EC) effects. After an acquisition phase in which neutral stimuli were related to positive or negative stimuli via instructions (Experiments 1 and 2a) or actual pairings (Experiment 2b), three different groups of participants were either informed that in the next phase the neutral stimuli would be presented without positive or negative stimuli (extinction instruction), that the neutral stimuli in the next phase would be paired with stimuli of the opposite valence than before (counterconditioning instruction), or received no further instructions. Afterwards, liking of the originally neutral stimuli was measured either with an evaluative rating (Experiment 1) or with an Implicit Association Test (IAT; Experiments 2a and 2b). EC was reduced in the counterconditioning condition of Experiment 1 and in the joint analysis of Experiments 2a and 2b. The extinction instruction led to a reduction of EC only in Experiment 1. Finally, whether the acquisition phase consisted of instructions about CS-US pairings (Experiment 2a) or the actual experience of CS-US pairings (Experiment 2b) did not significantly impact the observed changes in liking. Overall, our results suggest that similar mechanisms might mediate instruction- and experienced-based EC. Our results are in line with propositional models of EC but can be explained also by association formation models and dual process models of EC, provided that certain auxiliary assumptions are made.

Keywords: Evaluative conditioning, instructions, extinction, counterconditioning, propositional model

The Influence of Extinction and Counterconditioning Instructions on Evaluative Conditioning Effects

Evaluative conditioning (EC) is a change in the valence of a stimulus (conditioned stimulus or CS) that results from a previous pairing of the stimulus with another stimulus, the US (unconditioned stimulus) (e.g., De Houwer, 2007; Gast, Gawronski, & De Houwer, 2012; Levey & Martin, 1975). EC is considered to be an important way in which implicit and explicit evaluations can be changed. In order to learn more about this important phenomenon, EC researchers have tried to uncover the conditions under which it occurs and the mechanisms that mediate it (for reviews see De Houwer, Baeyens, & Field, 2005; De Houwer, Thomas, & Baeyens, 2001; Jones, Olson, & Fazio, 2010; for a meta-analysis see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010).

An important class of moderators that has been repeatedly studied in EC research are changes in the CS-US contingency. Examples for such changes in CS-US contingency are extinction or counterconditioning phases. In an extinction phase, CSs that were previously paired with positive or negative USs are presented alone, that is, without a US. In a counterconditioning phase, the participant continues to see CS-US pairings, but the valence of the US with which a particular CS is paired, is opposite to the valence of the US with which it was paired previously (e.g., a CS that was first paired with a positive US is paired with a negative US).

Extinction in particular has been studied extensively, although with mixed results. Most studies have shown that EC effects are resistant to the effects of an extinction phase: extinction trials did not significantly influence the size of the EC effect (Baeyens, Crombez, Van den Bergh, & Eelen, 1988; Blechert, Michael, Williams, Purkis, & Wilhelm, 2008; De Houwer, Baeyens, Vansteenwegen, & Eelen, 2000; Díaz, Ruiz, & Baeyens, 2005; Hermans,

Crombez, Vansteenwegen, Baeyens, & Eelen, 2003; Kerkhof et al., 2009; Vansteenwegen, Francken, Vervliet, De Clercq, & Eelen, 2006). Only a much smaller number of studies found that EC can be reduced by presenting extinction trials (Lipp, Mallan, Libera, & Tan, 2010; Lipp, Oughton, & LeLievre, 2003). A recent meta-analysis, however, confirmed that across studies, EC effects are smaller after than before an extinction procedure, although the EC effects after extinction are still substantial (Hofmann et al., 2010). This suggests that some of the studies in which an extinction phase was not found to influence EC might have suffered from a lack of power to detect a reduction of the EC effect (see also Lipp & Purkis, 2006, for a moderator that might influence whether extinction effects are found).

Only a few studies have investigated the effect of a counterconditioning procedure in EC. The results of these studies, however, are quite consistent and confirm that EC can be reduced by a counterconditioning phase (Baeyens, Eelen, Van den Bergh, & Crombez, 1989; Kerkhof, Vansteenwegen, Baeyens, & Hermans, 2011; Lipp et al., 2010).

In a prototypical EC study, the participant is presented with multiple stimulus pairings. Recently, however, it has been demonstrated that EC effects can also be found if the participant is merely instructed about the pairings and does not actually perceive them. De Houwer (2006) informed participants that nonwords such as “Bayram” or “Udibnon” (CSs) would be paired with positive or negative photos (USs). After reading these instructions, but without actually seeing the pairings, the participants performed an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) that provides an indirect measure of the valence of the stimuli. De Houwer showed that nonwords that were announced to be later paired with positive photos were evaluated more positively in an IAT than nonwords that were announced to be paired with negative photos. More recently, Gast and De Houwer (2012) showed that EC without actual pairings can also be found after instructions that only imply and do not explicitly mention the pairings. In one experiment, participants repeatedly

saw positive and negative USs that were accompanied by a grey square and a number that depended on whether the US was positive (e.g., the Number 1) or negative (e.g., the Number 2). Later on, participants were informed that the grey square covered one particular CS picture whenever the Number 1 was displayed and that it covered another CS picture whenever the Number 2 was displayed. This information implied that one CS co-occurred with a positive US whereas the other CS co-occurred with a negative US. In line with this information, the former CS was preferred over the latter one.

Showing that EC effects can be based not only on actually experienced pairings, but also on instructions about pairings is not only important in its own sake, but also for the information it gives on the mental processes that could underlie EC. Typically, three classes of EC models are distinguished: propositional models, association formation models, and dual process models. According to propositional models, all EC effects are due to the formation and validation of propositional knowledge about CS-US relations (De Houwer, 2009; Mitchell, De Houwer, & Lovibond, 2009a). To the degree that mere instructions about CS-US pairings and actual experience of CS-US pairings result in the same propositions about the CS-US relation, propositional models of EC predict comparable EC effects with both types of acquisition.

Association formation models, on the other hand, typically say little about the possible effect of instructions about CS-US pairings. According to these models, EC effects are based on the (automatic) formation of associations between the CS and the US or between the CS and an evaluative response to the US during experience of the CS-US pairings (e.g., Baeyens, Eelen, & Crombez, 1995; Jones, Fazio, & Olson, 2009). Association formation models typically emphasize the relevance of repeated direct experience of CS and US and state that conscious propositional knowledge about the pairings is not crucial for EC (e.g., Baeyens, Eelen, Crombez, & Van den Bergh, 1992; Baeyens et al., 1995; Smith and DeCoster, 2000;

Strack & Deutsch, 2004). Hence, on the basis of prototypical association formation models of EC, one would expect that mere instructions about CS-US pairings would not lead to the same effects as the actual experience of CS-US pairings. Although one can envisage variants of association formation models that do allow for EC via instructions (e.g., Field, 2006), finding important parallels between instruction-based and experience-based EC would put serious constraints on this class of models (i.e., limit the type of models that are plausible).

Finally, it has recently been proposed that EC might depend on both propositional and association formation processes (e.g., De Houwer, 2007; Gawronski & Bodenhausen, 2011). Like single-process propositional models of EC, such dual process models of EC can explain EC via instructions by attributing it to the formation and evaluation of propositional knowledge about CS-US relations. However, depending on when a dual process model postulates propositional and when associative processes to take place, it might predict differences between EC via instruction and EC via experience. Such differences would, for instance, emerge if association formation processes (a) operate under different conditions than propositional processes and (b) are involved only in EC via experience. Therefore, learning more about the similarities and differences between instruction-based and experience-based EC effects can also aid the development of dual process models of EC.

In his initial studies on instruction, De Houwer (2006) focused on the basic EC effect, that is, the effect of instructions about the presence of CS-US pairings on CS valence. An important next step is to examine the effect of instructions about procedures that have been shown to moderate EC effects. In the present studies, we examined whether EC effects are moderated by instructions about extinction and counterconditioning procedures. That is, rather than exposing participants to an extinction procedure (i.e., presenting CS-only trials after CS-US trials) or to a counterconditioning procedure (i.e., pairing a CS with a US of different valence than the US it was paired with during acquisition), we merely instructed

participants that they would be exposed to such phases. In order to test the generality of our findings, we investigated the effects of instructions about extinction and counterconditioning phases both on EC that resulted from instructions about CS-US pairings and on EC that resulted from actual CS-US pairings.

To the best of our knowledge, the effect of instructions about extinction and counterconditioning phases has so far been investigated in only one set of studies (Lipp et al., 2010). The authors presented counterconditioning or extinction phases either with or without instructions that announced the change in contingency before it actually took place. Evaluative ratings that were collected several times during each block gradually changed after the actual change of contingency. These changes were not influenced by additional instructions and did not occur at a point in time where only the instruction had been given but the change in contingency had not yet occurred. Hence, the data of Lipp et al. suggest that instructions about changes in contingencies have little or no effect on EC. Although speculative, it is possible that these null effects arose because of specific aspects of the procedure. For instance, counterconditioning instructions in the studies by Lipp and colleagues merely stated that the pairings in the next phase would be “reversed”. Participants therefore still had to infer what they would see in the next phase. It is possible that not all participants made the effort to figure this out before the start of the next phase. More generally, it is possible that the participants in Lipp et al.’s study did not pay much attention to the instructions.[1] In order to ensure that the participants in our studies did process the instructions thoroughly, we emphasized that participants had to remember the instructions later on in order to finish the experiment successfully.

Experiment 1

In Experiment 1, instructions about an extinction or counterconditioning phase were given after instructions about CS-US pairings. In this initial study, we opted for an instruction-based rather than an experienced-based acquisition phase because we considered it more likely that instructions about an extinction or counterconditioning phase were effective if they followed an acquisition phase of the same format. In Experiments 2a and 2b, however, the effects of extinction and counterconditioning instructions were tested both after an instructed acquisition phase (Experiment 2a) and after an experienced acquisition phase (Experiment 2b). Data from these experiments suggested that the type of acquisition does not seem to matter that much after all.

Participants were first told that they would see one type of product paired with positive photos and another type of product paired with negative photos. Afterwards and depending on condition, they were either informed that in a second phase the products would no longer be paired with photos (extinction condition), that in a second phase the products would be paired with photos of the opposite valence (counterconditioning condition), or they were not informed about a second phase (control condition). CS valence was in all conditions measured after the last set of instructions by means of valence ratings.

Please note that extinction and counter-conditioning effects are thus not tested in a pre-post design (i.e., a first rating before the extinction or counterconditioning instruction compared with a second rating after those instructions). We preferred a between-participants approach because the repeating of the rating phase itself might bias the results. More specifically, recent events (i.e., events that occurred just before the final ratings) are known to have more impact when CSs have to be rated repeatedly than when they have to be rated only once (e.g., Collins & Shanks, 2002; Lipp & Purkis, 2006; Matute, Vegas, & De Marez, 2002). We wanted to avoid such sequence effects and therefore opted for a single measurement between-participants design. Because participants were assigned randomly to the different

conditions, it is unlikely that an effect of condition is due to anything else than the only procedural difference between the conditions, that is, the nature of the instructions.

Method

Participants. Seventy-five students participated in this and an unrelated experiment in return for either four Euro or course credit. Two participants did not enter ratings. Three participants were accidentally assigned to the study after having participated in a related pilot study.[2] Therefore their data were dropped from the analysis but this did not alter the conclusions. The final sample consisted of 70 participants (14 men) who were randomly assigned to the conditions "control" (n = 24), "extinction" (n = 23), or "counterconditioning" (n = 23). Their ages ranged from 18 to 33 years (M = 21.33; SD = 2.97).

Materials. The stimuli used as CSs were two pictures of fictitious commercial products (Pleyers, Corneille, Luminet, & Yzerbyt, 2007), which had been successfully used in our lab before (Gast & De Houwer, 2012). Which of the two pictures (toothpaste, toilet paper) was in the first phase announced to be paired with positive photos (CS_{pos}) and which was announced to be paired with negative photos (CS_{neg}) was counterbalanced across participants.

Procedure. After participants had given informed consent, they were seated in front of a computer screen from which they received all instructions (see Appendix). Participants first read that they would participate in a learning experiment and that they should read and remember the instructions carefully. Subsequently, all participants received instructions that one of the CSs would be paired with positive USs and the other with negative USs. Afterwards, participants in the extinction condition received instructions about a second phase in which the CSs would be shown without USs. Participants in the counterconditioning condition received instructions about a second phase in which the CSs would be paired with USs of the opposite valence. Participants in the control condition received no further

instructions. As a way to encourage participants to process the instructions thoroughly, we told them that they needed to memorize the instructions in order to complete the task successfully.

After reading the instructions, participants were asked to rate their liking of each CS (see Appendix for instructions) on a scale ranging from -10 to +10 by clicking on a value with the computer mouse. CSs were presented in random order.

Finally, it was announced that memory for the instructions would be tested before the participant would go on to the learning phase. In the extinction and counterconditioning conditions, memory testing was done separately for the two phases, first for the instructions about the first and then for the instructions about the second phase. For each phase, both CSs were shown one-by-one in random order. The participant was asked to indicate whether it would be followed by (1) positive pictures, (2) negative pictures, (3) positive and negative pictures, or (4) would NOT be followed by positive or negative pictures.

Afterwards, participants were informed that the experiment was finished. They were debriefed about the purpose of the experiment, explaining that the announced learning phase would not follow anymore, and dismissed.

Design. The experiment has two main experimental factors: the valence of the US a CS was paired with in the first phase according to the instruction (valence: CS_{pos}, CS_{neg}; within) and the type of instruction a participant received about the second phase (instruction type: no instruction, extinction, counterconditioning; between). The assignment of product stimulus to valence condition (stimulus assignment: toothpaste is CS_{pos}, toilet paper is CS_{pos}; between) was counterbalanced across participants.

Results

Memory for Instructions. Sixteen participants (22.86 %) made at least one error when asked to indicate the instructed pairings of the relevant phases. Please see Table 1 for details.

EC effects. A three-way ANOVA with the factors valence (as assigned in the first phase: CS_{pos}, CS_{neg}; within), instruction type (control, extinction, counterconditioning; between), and stimulus assignment (stimulus assignment: toothpaste is CS_{pos}, toilet paper is CS_{pos}; between) showed a main effect of valence, $F(1,64) = 12.62$, $p < .001$, $\eta^2_{partial} = 0.16$. This indicates a preference for the product that according to the instruction would in the first phase be paired with positive photos over the product that would be paired with negative photos (see Figure 1 for descriptive results). There was also an interaction of valence and instruction type, $F(2,64) = 4.01$, $p = .023$, $\eta^2_{partial} = 0.11$. Contrast analyses (based on an ANOVA involving the EC effect score that we calculated by subtracting the rating of the CS_{neg} from the rating of the CS_{pos} and otherwise the same factors) showed that the EC effect in the counterconditioning condition was significantly smaller than the one in the control condition, $p = .015$. Also the EC effect in the extinction condition was significantly smaller than the EC effect in the control condition, $p = .020$. Simple t-tests showed that the difference between CS_{pos} and CS_{neg} (i.e., EC) was significant in the control condition, $t(23) = 4.00$, $p < .001$, $d = 0.82$, but neither in the extinction condition, $t(22) = 0.76$, $p = .46$, $d = 0.16$, nor in the counterconditioning condition, $t(22) = 1.06$, $p = .30$, $d = 0.22$. In addition, we observed an interaction of stimulus assignment and valence, $F(1,64) = 11.44$, $p = .001$, $\eta^2_{partial} = 0.15$, indicating a more pronounced EC effect if toothpaste was the CS_{pos} and toilet paper was the CS_{neg} than when this assignment was reversed. This interaction indicates a general preference for the toothpaste over the toilet paper.[3]

Discussion

After participants were instructed about a first phase in which CS-US pairings would be presented, we informed some of the participants about a second phase in which the CSs would not be paired with a US (extinction condition) or would be paired with a US of opposite valence than in Phase 1 (counterconditioning condition). Afterwards liking of the CS was assessed with a rating scale. Memory for the instructed pairings was assessed with a forced-choice task.

About 77% of participants correctly indicated all instructed pairings at the end of the experiment, suggesting that the majority of participants processed the instructions thoroughly. Most importantly, the instructions about Phase 2 had a significant influence on the liking or disliking of the stimuli. We found that the EC effects in both the counterconditioning and extinction condition were significantly reduced compared to the EC effect in a control condition and were clearly non-significant. Note, however, that the reduction in the extinction condition has to be interpreted with some caution because the extinction effect was not found in a smaller sample that only comprised participants with correct memory for the instructions (see Footnote 2).

In addition to providing the first demonstration of instructed extinction and instructed counterconditioning in EC, our results also provide a replication of the instructed EC effect that was first reported by De Houwer (2006). Moreover, our results go beyond this earlier finding in that we used standard evaluative ratings as the dependent variable rather than an implicit measure of valence (IAT). Our results thus attest to the generality of the findings of De Houwer.

Although the use of evaluative ratings in our study can thus be regarded as a strength, it is also a weakness. It is generally accepted that evaluative ratings are more susceptible to demand effects than implicit valence measures such as the IAT. For this reason, the present

demonstration of instructed EC is more likely to have been due to demand compliance than the original demonstration of De Houwer (2006). In the following studies, we therefore used an implicit measure of liking.

Experiments 2a and 2b

The goal of Experiments 2a and 2b was to replicate the findings on instruction-based extinction and counterconditioning with an implicit measure that is less sensitive to demand compliance than standard evaluative ratings. We chose the IAT for this purpose. Although IAT effects can be controlled consciously under some conditions (e.g., De Houwer, Beckers, & Moors, 2007; Steffens, 2004), they are clearly more difficult to control than valence ratings. Furthermore, for the current purpose the IAT is more suitable than evaluative-priming-based measures, which for reasons of lower reliability (e.g., De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009) are not ideal for comparing EC effects of potentially differing sizes and which are also not immune to conscious control (e.g., Teige-Mocigemba & Klauer, in press).

A second goal of Experiment 2 was to investigate whether instruction-based extinction and counterconditioning phases lead to the same pattern of results if the initial acquisition phase is not instructed but actually experienced. The acquisition phase of Experiment 2b therefore involves actual presentations of CS-US pairs, while the acquisition phase of Experiment 2a is, like the one in Experiment 1, based on instructions. Finally, to test the generality of our findings, in Experiments 2a and 2b different stimuli were used than in Experiment 1. As in Experiment1, we used a between-participants comparison to estimate extinction and counterconditioning effects. In the present study, such a design was necessary because the results of the IAT are known to change as a function of previous experience with the IAT (Nosek, Banaji, & Greenwald, 2002). Therefore, pre-post differences in IAT effects

might be due not to the effect of instructions but merely to the effect of experience with the IAT.

Method

Participants. Eighty-five students (21 men) were paid eight Euros for their participation in Experiment 2a and an unrelated experiment. Twenty-eight of these were assigned to the condition "control", 29 to the condition "extinction", and 28 to the condition "counterconditioning". Ninety-one students (33 men) participated in Experiment 2b in exchange for course credit or four Euros. Twenty-three of these were assigned to the condition "control", 30 to the condition "extinction", and 38 to the condition "counterconditioning".[4]

Materials. The nonwords "UDIBNON" and "ENANWAL" served as CSs. Which of these was used as CS_{pos} and which as CS_{neg} was counterbalanced across participants. During the IAT, these nonwords were presented as targets in four different fonts (Algerian, Arial Black, Impact, and Comic Sans MS), size 34. The Dutch words for 'SINCERE', 'HAPPY', 'HONEST', 'NICE', 'MEAN', 'BRUTAL', 'AGRESSIVE' and 'FAKE' were used as positive and negative attribute stimuli in the IAT. These words were presented in size 34, font Arial Black. Ten positive (1440, 1710, 2209, 2216, 2310, 2340, 2530, 5621, 5779, 8540) and ten negative (1280, 2120, 2490, 2710, 2800, 6940, 9001, 9040, 9140, 9300) IAPS pictures (Lang, Bradley, & Cuthbert 2008) were used as USs in the acquisition phase of Experiment 2b.

Procedure. The first phase of Experiment 2a consisted of instructions about upcoming CS-US pairings that were similar to those used in the first phase of Experiment 1 (see Appendix for instructions).

The first phase of Experiment 2b consisted of actual CS-US pairings. Participants first read a general instruction that announced that positive and negative photos would be shown, always preceded by a nonword. Participants were asked to watch the stimuli attentively (see Appendix for instructions). Then the stimulus pairings were presented. Each CS was shown ten times. The CS_{pos} was always followed by a positive photo whereas the CS_{neg} was always followed by a negative photo. Ten different positive and ten different negative pictures were presented as USs, so that each of these appeared once. Each trial started with a blank screen for 200 ms, and then the CS was presented for 2000 ms. After a stimulus interval of 500 ms during which the screen was blank, the US appeared and stayed on the screen for 5000 ms. The trial ended with a blank screen for 1800 ms.

The instructions about the second phase of Experiments 2a and 2b informed participants in the extinction condition that the CSs (the nonwords) would be presented alone in the second phase. Participants in the counterconditioning condition were informed about a reversed CS-US assignment (see Appendix for the instructions). In the control condition, no instructions were given.

After the learning phase(s), participants were asked to complete a reaction time task. If the last phase was instruction-based (all conditions of Experiment 2a and the extinction and counterconditioning conditions of Experiment 2b), participants were told that they had to do the reaction time task first and were asked to keep the instructions in memory. Before the IAT started, participants were informed that words would appear on the screen one-by-one. There were four types of words: (1) positive words (e.g., happy), (2) negative words (e.g., fake), (3) the word UDIBNON, (4) the word ENANWAL. Participants were told that depending on the type of word, they had to press the left (Q) or right key (M). They were informed that the assignment of word types to keys would differ from phase to phase and that this would be indicated by the appearance of labels assigned to the left key in the upper left corner of the

screen and labels assigned to the right key in the upper right corner of the screen. Participants were asked to respond as quickly as possible without making too many errors.

The IAT consisted of the following blocks: (1) A practice block of 32 trials during which the two nonwords were presented 16 times each, (2) a practice block of 16 trials in which the four positive and four negative adjectives were presented twice each, (3) and (4) two test blocks of 32 trials in which each of the two nonwords was presented eight times and each of the affective words was presented twice, (5) a second practice block of 32 trials of only nonwords with reversed response assignments, and (6) and (7) two test blocks of 32 trials with reversed assignment of the nonwords in which each of the two nonwords was presented eight times and each of the affective words was presented twice. The trials within a block were presented in random order. Before each block, categories and response assignments relevant in this block were announced. During each block, the relevant category labels (“POSITIVE”, “NEGATIVE”, “UDIBNON”, “ENANWAL”) were indicated in the left or right upper corner of the screen, depending on the current response assignment.

Each IAT trial started with the presentation of a word in the center of the screen. When a correct response was given, the word disappeared. When an incorrect response was given, the word was replaced by a red X, which stayed on the screen for 400 ms. In both cases, the next word appeared after an inter-trial-interval of 400 ms.

Finally, participants were asked several questions. First, their memory for the instructed pairings of each of the CSs was tested with a similar question as in Experiment 1. After each of these questions, participants were asked to indicate how certain they were about their response. Next, participants were asked to indicate on a rating scale ranging from 1 to 9 how pleasant they found the nonwords (CSs). Next, but only for Experiment 2b, participants were asked to rate the presented USs. Also only for Experiment 2b, participants were asked an

open question about whether there was a regularity in the order in which photos and words were presented in the first phase. Finally, participants from Experiment 2b were – with a similar question as for the instructed pairs – also asked for each CS by which type of photo it was followed in the first (experienced) phase. Also here, participants were asked to indicate their certainty. Most of these questions were added for exploratory reasons and will not be discussed further.

Design. In order to examine the impact of actual vs. instructed CS-US pairings in Phase 1, the data from Experiment 2a (instructed CS-US pairings in Phase 1) and Experiment 2b (actual CS-US pairings in Phase 1) were analyzed together, with experiment (2a or 2b) treated as a between-subjects factor. The main experimental factor in the overall analysis was the type of instruction a participant received in Phase 2 (instruction type: control, extinction, counterconditioning; between). Assignment of nonword to valence condition (stimulus assignment: UDIBNON is CS_{pos}, ENANWAL is CS_{pos}), IAT order (congruent or incongruent blocks first), and assignment of the positive and negative valence categories to the right and left hands in the IAT (hand assignment: positive right, negative right) were counterbalanced across participants.

Results

Memory for Instructed and Experienced Pairings. In Experiment 2a, seventeen participants (20.00 %) made at least one error when asked to indicate the instructed pairings. In Experiment 2b, twenty-three participants (25.27 %) made at least one error when asked to indicate the experienced or instructed pairings. Please see Table 1 for details.

EC effects as measured with the IAT. The IAT data were prepared following one of the recommended scoring algorithms (D_4; see Greenwald, Nosek, & Banaji, 2003). The *D*-measure is based on the response time difference between incongruent and congruent blocks

divided by the relevant standard deviation. We consider blocks as congruent in which a CS was assigned to the same key as the valence of the US it was paired with in the first phase (actually or according to instructions). Blocks with reversed assignment are considered as incongruent. Hence, larger *D*-values indicate a more pronounced EC effect in line with the instructed or experienced pairings from the first phase. We conducted an ANOVA with the factors instruction type, experiment, stimulus assignment, IAT order, and hand assignment (see Figure 2 for descriptive results). There was a significant main effect of instruction type, $F(2, 128), = 5.29$, $p = .006$, $\eta^2_{partial} = .076$. Contrasts showed that the control and extinction conditions did not differ significantly from each other, $p = .764$. The EC effect in the counterconditioning condition, however, differed significantly from that in the control condition, $p = .012$. Simple t-tests showed that the D4-value was significantly above zero in the control condition, $t(50) = 7.74$, $p < .001$, $d = 1.08$, in the extinction condition, $t(58) = 6.14$, $p < .001$, $d = 0.80$, and also in the counterconditioning condition, $t(65) = 3.15$, $p = .002$, $d = 0.39$. The ANOVA also revealed some less relevant effects. First, we observed a significant main effect of stimulus assignment, $F(1,128) = 5.49$, $p = .021$, $\eta^2_{partial} = 0.041$, indicating a more pronounced EC effect when "ENANWAL" was the CS_{pos}, which indicates a general preference for "ENANWAL" over "UDIBNON". Second, there was a significant main effect of IAT order, indicating a more pronounced IAT effect when the congruent block was worked on first, $F(1,128) = 23.18$, $p < .001$, $\eta^2_{partial} = 0.153$, which is a common finding with the IAT. Neither the main effect of experiment nor its interaction with instruction type was significant, both *F*'s < 1. Even though there were no effects of experiment, we also performed analyses separately for Experiments 2a and 2b. The data of Experiment 2a (instructed CS-US pairings as first phase) did not show a significant main effect of instruction type, $F(2, 61), = 1.26$, $p = .291$, $\eta^2_{partial} = .040$. In Experiment 2b (actual CS-US pairings as first phase), however, we did find a main effect of instruction type, $F(2, 67), = 5.62$, $p = .006$, $\eta^2_{partial} = .144$. Contrast

analyses on the data of Experiment 2b showed that the control and extinction conditions did not differ significantly from each other, $p = .407$. The difference between the control and the counterconditioning condition was significant, $p = .036$.[5]

Discussion

Eighty percent of participants in Experiment 2a and about 75% of participants in Experiment 2b correctly indicated all instructed pairings at the end of the experiment, suggesting that the majority of participants had processed the instructions thoroughly. The instructed counterconditioning effect that was for the first time observed in Experiment 1 was replicated. Informing participants that a phase of reversed pairings would follow as a second phase decreased the EC effect. However, unlike to what was the case in Experiment 1, the EC effect in the counterconditioning condition did not disappear completely. Also contrary to the results of Experiment 1, the extinction instruction did not have a significant impact on EC in Experiments 2a and 2b. The findings were not significantly moderated by whether the first phase was instruction- or experience-based.

The results from Experiments 2a and 2b once again replicated the instructed EC effect that was first observed by De Houwer (2006), showing that instructions about pairings can lead to significant EC effects. Importantly, the EC effects after instructions about pairings (Experiment 2a) or actually experienced pairings (Experiment 2b) did not differ in size. It is informative that in an additional comparison that only involved the control conditions of Experiments 2a and 2b (which only had the first acquisition phase), the instructed EC effects also did not differ in magnitude, $t(50) = 0.10$, $p = .92$. This suggests that actually experiencing the pairings does not lead to stronger EC effects than being instructed about them. Please note, however, that these comparisons have to be interpreted with the caution required for between-experiments-comparisons.

General Discussion

In three experiments, we investigated the influence of instructions about changes in CS-US contingencies on EC. After an acquisition phase that was either instructed or experienced, participants in different experimental conditions were either instructed about a second phase in which the CSs would not be followed by USs anymore (extinction conditions) or a second phase in which the CSs would be paired with USs of the opposite valence as before (counterconditioning conditions). In the control conditions, participants were not instructed about a second phase.

A first important finding is that instructions about reversed contingencies (counterconditioning) consistently led to a substantial reduction of the EC effect. The EC effect in the counterconditioning condition of Experiment 1 was reduced to non-significance. In Experiment 2, the EC effect in the counterconditioning condition was reduced but still significant.

The second important set of findings concerns the effect of an instruction-based extinction phase. Here the results were mixed. While the results of Experiment 1 showed decreased EC after an instructed extinction phase, the results of Experiments 2 showed an EC effect of at least equal size in the extinction as in the control group. The significant extinction effect in Experiment 1 should be interpreted cautiously, however, given that it was not significant for the subset of participants who correctly remembered all instructions.

If we do want to take the significant extinction effect in Experiment 1 seriously, it could on the one hand be compared to the non-significant extinction effect on the IAT scores in Experiment 2 and on the other hand to the non-significant extinction effects that were reported in a large number of studies. A possible explanation for the discrepancy in the results of Experiments 1 and 2 lies in the valence measure used. Earlier research suggests that liking

as assessed with implicit measures (i.e., implicit evaluation) is less easy to change than liking that is assessed with rating measures (i.e., explicit evaluation; see Gregg, Seibt, & Banaji, 2006; Peters & Gawronski, 2011). Hence, information about changes in stimulus pairings might also be less likely to influence liking assessed with implicit measures than liking assessed with rating measures. This would explain why instructed extinction was observed only in Experiment 1 (in which a rating measure was used) but not in Experiments 2a and 2b (in which an implicit measure was used). This reasoning is also congruent with the fact that counterconditioning instructions eliminated the EC effect in Experiment 1 but only reduced it in Experiment 2.

Regarding the question of why Experiment 1 showed significant extinction while the majority of studies with experienced-based extinction trials did not, one first has to consider that some EC studies did show significant extinction due to experiencing CS-only trials. In fact, a recent meta-analysis revealed that EC is sensitive to extinction (Hofmann et al., 2010). Hence it is not clear to what degree the significant instruction-based extinction effect actually deviates from what is known about experience-based extinction. Nevertheless, one could speculate whether instructing a participant that the CS will be shown alone is more effective than actually presenting CS-only trials to the participant. Extinction trials are always shown at the end of the learning phase. It is possible that participants who first experience a series of acquisition trials followed by a series of extinction trials might get bored over the course of trials and pay less attention to the extinction trials at the end of conditioning phase. It is also possible that participants actually consider the extinction trials as less interesting (for example because no valent stimuli appear) and therefore pay less attention. Instruction-based extinction is less likely to suffer from this loss of attention. First, instructions about trials take less time than the trials themselves; instruction trials might therefore suffer less from a gradual decrease of attention. Second, the fact that the second phase is mentioned in the same

way as the first phase suggests that it is important. For the reasons mentioned above, however, these ideas should be treated as mere speculation.

A third important result is that we replicated instructed EC using both a standard evaluative rating measure and an IAT measure. Furthermore, comparing Experiments 2a and 2b allowed us to compare instruction-based and experienced-based EC for the first time. Using an IAT measure, we found no difference between the experiments, neither with regard to the effect of experience vs. instruction regarding the initial CS-US pairings (i.e., the EC effects did not differ in the control conditions of Experiments 2a and 2b), nor with regard to the effect of extinction and counterconditioning instructions.

One of the reasons why we set out to test the effects of instructed procedures on EC was that similarities and differences between the effects of instructed and experienced procedures could inform us about the mechanisms that mediate EC. Our initial results do suggest that instructions about stimulus contingencies and the actual experience of these stimulus contingencies have quite similar effects. This holds both for initial instructions about CS-US contingencies and for subsequent instructions about changes in the CS-US contingencies.

So what do these similarities tell us about the mechanisms that mediate EC? In our opinion, the observed similarities are in line with propositional models of associative learning (De Houwer, 2009; Mitchell et al., 2009a). These models postulate that EC effects depend on the acquisition and validation of propositional knowledge about the stimulus pairings.[6] Such knowledge can be acquired both by observing the actual pairings and by being informed about them. In fact propositional models predict no difference between EC effects due to instructions and due to experience, provided that the acquired propositional knowledge about stimulus relations is similar in content and in how valid the knowledge is considered to be.

Note that, irrespective of whether EC is based on instructions or experience, propositional models can explain not only acquisition effects, but also effects of extinction and counterconditioning. According to these models, one could assume that in addition to the proposition about the first phase (e.g., “this product co-occurs with positive pictures”), a proposition is formed about the stimulus relations in the second phase (e.g., “this product co-occurs with negative pictures”) that counteracts the effect of the first proposition. Alternatively, one could assume that a proposition is formed that is thought to apply to both phases (e.g., “this product is sometimes paired with positive and sometimes with negative pictures”). Although it is not entirely clear how (e.g. involving which further processes) propositions lead to changes in liking (see Mitchell et al., 2009b, for a discussion of this issue), such changes in propositions are likely to lead to changes in liking. Propositional models might also account for why counterconditioning (instructions) leads more reliably to a change in valence than extinction (instructions). One could, for instance, argue that propositions formed after counterconditioning (instructions) (e.g., “this product is sometimes followed by positive pictures and sometimes by negative pictures”) are more likely to change liking than propositions formed after extinction (instructions) (e.g., “this product is sometimes followed by positive pictures and sometimes presented alone”).

How do our results relate to association formation models? As we indicated in the introduction, these models typically emphasize the relevance of repeated direct experience (e.g., Baeyens et al., 1992, 1995; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Simple association formation models, therefore, have difficulties in explaining any effect of instructions on the EC effect, independent of whether it informs a participant about contingencies in a second or in a first phase. Recent interpretations of associative models (e.g., Field, 2006), however, do allow for conditioning via instructions. For instance, it has been argued that an instruction about stimulus pairings (e.g., “product A is followed by

positive pictures") itself presents participants with a stimulus pairing (e.g., between product A and the word "positive" or between the mental representation of product A and the mental representation of positive pictures; see Field, 2006). Such arguments allow association formation models to explain the basic finding that instructions about pairings can lead to EC effects (De Houwer, 2006). It could even explain that counterconditioning instructions have an effect (e.g., because they result in new associations involving USs of an opposite valence) or that extinction instructions have an effect (e.g., because they weaken or modulate the original association). However, it is important to realize that association formation models can deal with these effects only if a single pairing of words in an instruction can lead to the formation of associations in much the same way as actual CS-US pairings. As such, our data heavily constrain association formation models.

As indicated in the introduction, dual process models can also account for instruction-based EC. Let us consider the well-known associative-propositional evaluation (APE) model of Gawronski and Bodenhausen (2011). It postulates that EC can in principle result either from association formation or from propositional processes. Hence, it is possible to argue that experience can lead to EC via the formation of associations whereas instructions can lead to EC via the formation of propositions. However, the APE model also allows for propositional processes to result in the formation of associations. As a result, the impact of instructions on liking could sometimes also be mediated by association formation. Likewise, once associations have been formed in memory, they can give rise to propositions. Given this high level of interactivity between the formation of associations and propositions, it is not always straightforward to determine when instruction-based and experienced-based EC will be similar and when they will differ. Nevertheless, the fact that we found few differences between both suggests that if EC effects are indeed based on the joint influence of associations and propositions, these are influenced by instructions in much the same way as

by actual experience. This of course raises the question of why both associations and propositions are needed in order to account for EC. Based on the argument of parsimony, we thus believe that our results fit best with a single-process propositional model of EC. Nevertheless, it remains important to continue looking for possible differences between instruction-based and experienced-based EC because these could provide important information about whether it is necessary to postulate multiple processes as sources of EC and, if so, how these sources interact.

Whereas we found clear effects of counterconditioning instructions on EC, Lipp et al. (2010) failed to find such effects. The two sets of studies differ in several respects, such as the type of stimuli used and the exact timing parameters of the experienced conditioning trials, which makes a discussion about the source of the differences very speculative. An important difference might, however, be the wording of the instructions. In our studies, instructions were very explicit in stating with which type of photo a CS would be paired in the second phase (e.g., “There is an important difference between the first and the second phase: If you see a photo of this product, a positive photo will appear”). In the study of Lipp and colleagues, on the other hand, less specific instructions were given (i.e., “IMPORTANT MESSAGE, The pairing of shapes and faces, will now be reversed”) that required participants to infer the nature of change themselves. Moreover, whereas we strongly encouraged participants to thoroughly process the instructions, there was less incentive for the participants in Lipp et al.’s study to do so.

Another issue that we would like to discuss is the issue of demand compliance. Demand compliance in evaluative conditioning can arise if (a) participants have strong beliefs about the experimental hypothesis, in this case that the CSs should change in valence depending on which US they were paired with (*demand awareness*), (b) participants know which CS was paired with which US (*contingency awareness*; for the distinction of demand

awareness and contingency awareness, see Field, 2000), (c) participants are *motivated* to show behavior that is in line with the hypothesis, and (d) participants *can control* their behavior in such a way that the observed responses are in line with the perceived hypothesis.

Considering these conditions for demand compliance, it is likely that the probability of demand compliance is higher for EC based on instructions than for EC based on experienced pairings. First, contingency awareness is typically high after instructing participants explicitly about the pairings. Second, it is possible that through the instructions participants become aware not only of the pairings themselves, but also of the fact that the pairings are relevant for the experimental hypothesis, which might increase the chance for becoming demand aware. Therefore, the use of implicit measures (which are more difficult to control than explicit measures) is particularly important when investigating instructed EC. However, also the IAT and other implicit measures are controllable under some conditions (Bar-Anan & Nosek, 2012; De Houwer et al., 2007; Steffens, 2004; Teige-Mocigemba & Klauer, in press). Demand compliance therefore always remains an alternative hypothesis for EC effects even when using implicit measures. Nevertheless most researchers would agree that implicit measures are clearly more difficult to control than explicit ratings. In addition to their difficulty to control, implicit measures might also decrease the impact of demand compliance by obscuring the experimental hypothesis (it is less obvious that the researcher is interested in the valence of the CS). In sum, it is important to realize that (a) demand compliance can take place only if several conditions are met and (b) implicit measures in several ways reduce the probability that these conditions are met. Hence, although it is difficult to ever exclude the possibility of demand compliance completely, its impact in studies should not be overestimated, especially when implicit measures are used.

A final point that we should at least shortly comment on is the question whether instructed EC is actually EC. EC is typically defined as a valence change in a stimulus that is

due to pairing the stimulus with another stimulus (e.g., De Houwer, 2007; Gast et al., 2012). At first sight, this definition does not seem to apply to instruction-based EC because an instruction about a stimulus pairing is not the same as a real repeated stimulus pairing. It has been argued, however, that also instructions can be seen as either involving actual stimulus pairings or as referring to actual stimulus pairings (see De Houwer, Barnes, Holmes, & Moors, in press; Field, 2006; Gast et al., 2012 for a discussion of this issue). Although interesting from a meta-theoretical point of view, the question of whether instruction-based EC is EC is in the current context merely a terminological issue. Independent of whether the current results qualify as EC, they do give new information about the determinants of stimulus preferences in general and about EC specifically. More specifically, by comparing changes in liking based on actual pairings with changes in liking based on instructions about pairings, we learn more about how both actual pairings and instructions about pairings influence liking.

To summarize, in three studies we investigated the impact of instructions about stimulus pairings on EC. In line with earlier findings (De Houwer, 2006), we observed that instructions about upcoming CS-US pairings gave rise to EC effects. Instruction-based counterconditioning (informing participants that the pairings would be reversed in a second phase) consistently led to a decrease in the EC effect. Instruction-based extinction (informing participants that the CSs would be presented alone in a second phase) reduced EC effects in only one of the studies. The overall pattern of results is similar to the findings reported after experience-based acquisition, counterconditioning, and extinction procedures. This surprising similarity might suggest that experience-based and instruction based EC are due to similar mental processes. We argued that these findings fit well with propositional models of EC. Our results do, however, diverge from those of Lipp et al. (2010). Future studies therefore need to focus on the boundary conditions of instruction-based extinction and counterconditioning effects in EC.

Acknowledgements

We thank Baptist Liefooghe, Sarah Opsomer, and Dorit Wenke for their contribution in planning Experiments 2a and 2b and Sarah Opsomer for conducting Experiments 2a and 2b.

References

Baeyens, F., Crombez, G., Vandenbergh, O., & Eelen, P. (1988). Once in contact always in contact – evaluative conditioning is resistant to extinction. *Advances in Behaviour Research and Therapy, 10,* 179-199. doi: 10.1016/0146-6402(88)90014-8

Baeyens, F., Eelen, P., & Crombez, G. (1995). Pavlovian Associations are Forever: On classical Conditioning and Extinction. *Journal of Psychophysiology, 9,* 127-141. Retrieved from http://www.hogrefe.com/index.php?mod=journals&action=1&id=19

Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy, 30,* 133-142. Retrieved from http://www.elsevier.com/wps/find/journaldescription.cws_home/265/description#description

Baeyens, F., Eelen, P., Vandenbergh, O., & Crombez, G. (1989). Acquired affective evaluative value – conservative but not unchangeable. *Behaviour Research and Therapy, 27,* 279-287. doi: 10.1016/0005-7967(89)90047-8

Blechert, J., Michael, T., Williams, S. L., Purkis, H. M., & Wilhelm, F. H. (2008). When two paradigms meet: Does evaluative learning extinguish in differential fear conditioning? *Learning and Motivation, 39,* 58-70. doi: 10.1016/j.lmot.2007.03.003

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38,* 1193-1207. doi 10.1177/0146167212446835.

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition, 30,* 1138-1147. doi: 10.3758/BF03194331.

Cook, S. W., & Harris, R. E. (1937). The verbal conditioning of the galvanic skin reflex. *Journal of Experimental Psychology, 21,* 202-210. doi:10.1037/h0063197

De Houwer, J. (2006). Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37,* 176-187.doi: 10.1016/j.lmot.2005.12.002

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10,* 230-241. Retrieved from http://www.ucm.es/info/psi/docs/journal/

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37,* 1-20. doi:10.3758/LB.37.1.1

De Houwer, J., Baeyens, F., & Field, A. P. (2005). Associative learning of likes and dislikes. Some current controversies and possible ways forward. *Cognition and Emotion, 19,* 161-174. doi: 10.1080/02699930441000265

De Houwer, J., Baeyens, F., Vansteenwegen, D., & Eelen, P. (2000). Evaluative conditioning in the picture-picture paradigm with random assignment of conditioned stimuli to unconditioned stimuli. *Journal of Experimental Psychology-Animal Behavior Processes, 26,* 237-242. doi: 10.1037/0097-7403.26.2.237

De Houwer, J., Barnes-Holmes, D., & Moors, A. (in press). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review.*

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology, 43,* 972-978. doi: 10.1016/j.jesp.2006.10.007

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit Measures: A Normative Analysis and Review. *Psychological Bulletin, 135,* 347-368. doi: 10.1037/a0014211

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127,* 853-869. doi: 10.1037//0033-2909.127.6.853

Diaz, E., Ruiz, G., & Baeyens, F. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cognition & Emotion, 19,* 245-268. doi: 10.1080/02699930441000300

Ebert, I. D., Steffens, M. C., von Stülpnagel, R., & Jelenec, P. (2009). How to like yourself better, or chocolate, less: Changing implicit attitudes with one IAT task. *Journal of Experimental Social Psychology, 45, 1098-1104.* doi: 10.1016/j.jesp.2009.06.008

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review, 26,* 857-875. doi:10.1016/j.cpr.2005.05.010

Gast, A. & De Houwer, J. (2012). Evaluative conditioning without directly experienced pairings of the conditioned and the unconditioned stimuli. *The Quarterly Journal of Experimental Psychology, 65,* 1657-1674. doi: 10.1080/17470218.2012.665061

Gast, A., Gawronski, B., & De Houwer, J. (2012). Evaluative Conditioning: Recent Developments and Future Directions. *Learning and Motivation, 43,* 79-88. doi: 10.1016/j.lmot.2012.06.004

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127. doi:10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus Contextualization in Automatic Evaluation. *Journal of Experimental Psychology-General, 139,* 683-701. doi: 10.1037/a0020315

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G, Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197-216. doi:10.1037/0022-3514.85.2.197

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90,* 1-20. doi: 10.1037/0022-3514.90.1.1

Hermans, D., Crombez, G., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2003). Expectancy-learning and evaluative learning in human classical conditioning: Differential effects of extinction. In P. L. Gower (Ed.), *Psychology of Fear* (pp. 133-156). New York: Nova Science Publishers, Inc.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136,* 390-421. doi:10.1037/a0018916

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of Personality and Social Psychology, 96,* 933-948. doi: 10.1037/a0014747

Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The "how" question. *Advances in Experimental Social Psychology, 43,* 205-255. doi: 10.1016/S0065-2601(10)43005-1

Kerkhof, I., Goesaert, E., Dirikx, T., Vansteenwegen, D., Baeyens, F., D'Hooge, R., et al. (2009). Assessing valence indirectly and online. *Cognition & Emotion, 23,* 1615-1629. doi: 10.1080/02699930802469239

Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning An Effective Technique for Changing Conditioned Preferences. *Experimental Psychology, 58,* 31-38. doi: 10.1027/1618-3169/a000063

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *The international affective picture system (IAPS): Technical manual and affective ratings*. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Levey, A. B., & Martin, I. (1975). Classical conditioning of human 'evaluative' responses. *Behaviour Research and Therapy, 4,* 205-207. doi: 10.1016/0005-7967(75)90026-1

Lipp, O. V., Mallan, K. M., Libera, M., & Tan, M. S. (2010). The effects of verbal instruction on affective and expectancy learning. *Behaviour Research and Therapy, 48,* 203-209. doi: 10.1016/j.brat.2009.11.002

Lipp, O. V., Oughton, N., & LeLievre, J. (2003). Evaluative learning in human Pavlovian conditioning: Extinct, but still there? *Learning and Motivation, 34,* 219-239. doi: 10.1016/s0023-9690(03)00011-0

Lipp, O. V., & Purkis, H. M. (2006). The effects of assessment type on verbal ratings of conditional stimulus valence and contingency judgments: Implications for the extinction of evaluative learning. *Journal of Experimental Psychology-Animal Behavior Processes, 32,* 431-440. doi: 10.1037/0097-7403.32.4.431

Matute, H., Vegas, S., & De Marez, P.-J. (2002). Flexible use of recent information in causal and predictive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 714–725. doi: 10.1037//0278-7393.28.4.714

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009a). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*, 183-198. doi:10.1017/S0140525X09000855

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009b). The propositional nature of human associative learning: Response to commentaries. *Behavioral and Brain Sciences, 32*, 230-246. doi: 10.1017/S0140525X09001186

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics, 6,* 101–115. doi: 10.1037//1089-2699.6.1.101

Peters, K. R., Gawronski, B. (2011). Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations. *Personality and Social Psychology Bullettin, 37*, 557-569. doi: 10.1177/0146167211400423.

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis)liking: Item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 130-144. doi:10.1037/0278-7393.33.1.130

Smith, E. R., & DeCoster, J. (2000). Dual process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131. doi:10.1207/S15327957PSPR0402_01

Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*, 165-179. doi: 10.1027/1618-3169.51.3.165

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review, 8*, 220-247. doi: 10.1207/s15327957pspr0803_1

Teige-Mocigemba, S. & Klauer, K. C. (in press). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition & Emotion.*

Vansteenwegen, D., Francken, G., Vervliet, B., De Clercq, A., & Eelen, P. (2006). Resistance to extinction in evaluative conditioning. *Journal of Experimental Psychology-Animal Behavior Processes*, *32*, 71-79. doi: 10.1037/0097-7403.32.1.71

Table 1

Numbers of participants who incorrectly indicated pairings for a specified phase by experiment and condition. Total numbers of participants in condition in brackets.

	Experiment 1		
	Control	Extinction	Counter
Phase 1	2(24)	3(23)	2(23)
Phase 2	-	8(23)	3(23)
Total	2(24)	9(23)	5(23)
	Experiment 2a		
	Control	Extinction	Counter
Phase 1	2(28)	4(29)	2(28)
Phase 2	-	8(29)	5(28)
Total	2(28)	10(29)	5(28)
	Experiment 2b		
	Control	Extinction	Counter
Phase 1	1(23)	1(30)	5(38)
Phase 2	-	6(30)	13(38)
Total	1(23)	6(30)	16(38)

Figure 1

Mean evaluative ratings of the entire participant sample from Experiment 1 for the factors valence and instruction type (marginal means). Error bars represent standard errors.

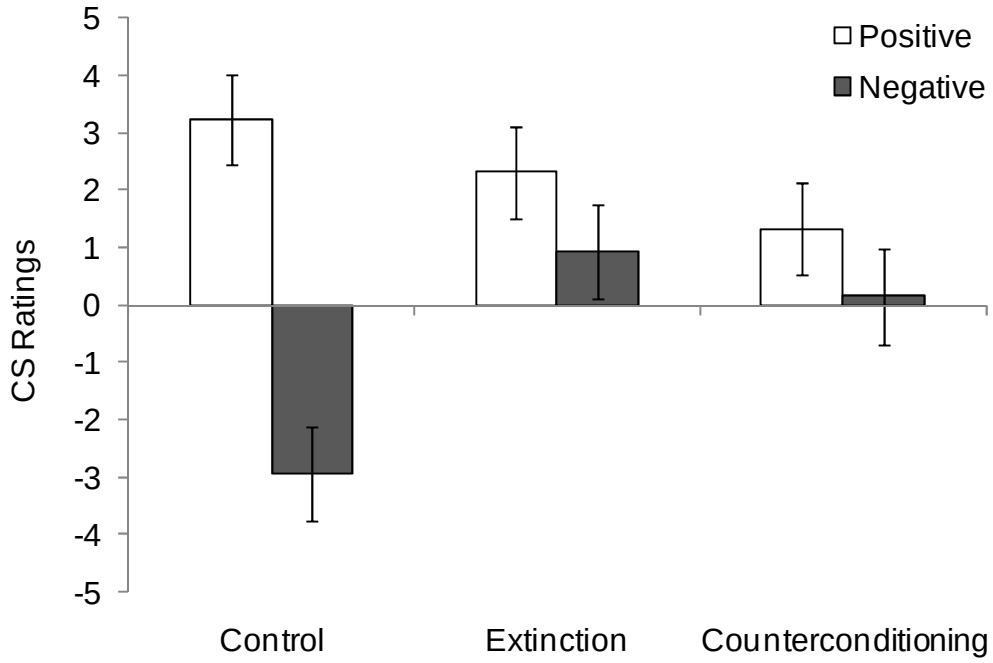
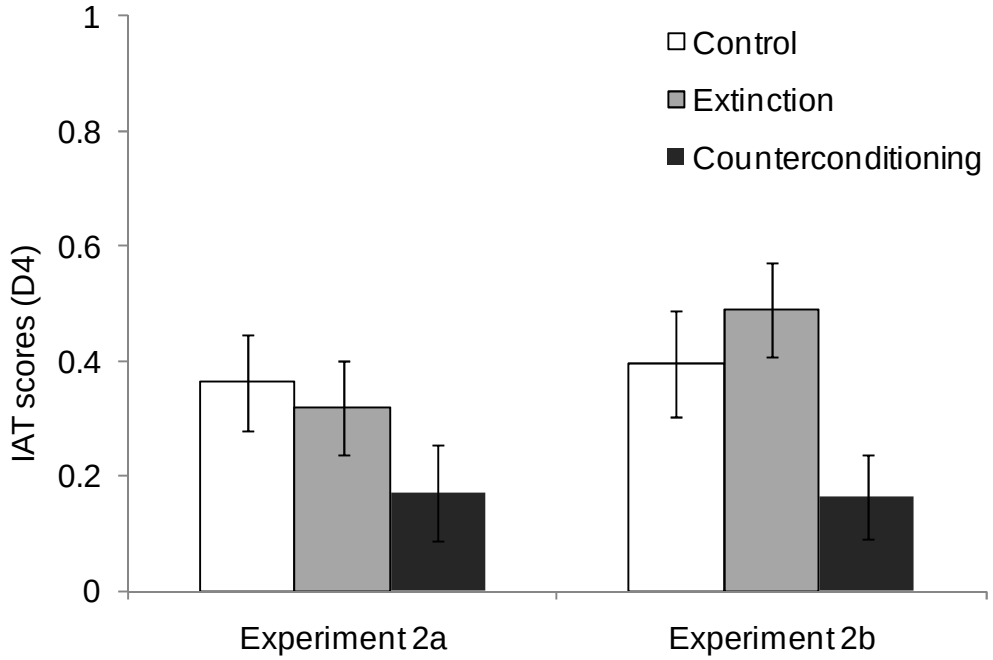


Figure 2

Mean D4 scores of the entire sample from Experiments 2a and 2b for the conditions of the factor instruction type (marginal means). Error bars represent standard errors.



Appendix: Instructions

(Translated from Dutch; comments in square brackets)

Experiment 1

Instructions Phase 1, all conditions. In a first phase of this learning experiment [control condition: “in the learning phase”] you will see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of maimed bodies). Each photo will be preceded by a photo of a product that indicates which type of photo (positive or negative) will appear.

If you see a photo of this product [display of Picture 1 (toilet paper) or Picture 2 (toothpaste), depending on counterbalancing condition] a positive photo will appear.

If you see a photo of this product [display of Picture 2 or Picture 1, depending on counterbalancing condition] a negative photo will appear.

It is very important that you now already remember which product goes together with which type of photo (positive or negative) [only in extinction and counterconditioning condition: “in this first phase”]. You will definitely need this information to finish the task successfully. This information will not be presented again, so remember well which product goes together with which type of photo [in this first phase].

Instructions Phase 2, extinction condition. After the first phase, follows a second phase.

[bold] Watch out: There is an important difference between the first and the second phase:

During the second phase of the learning experiment, you will only see the photos of the products.

[display of Picture 1 and Picture 2]

The photos of the products will during the second phase NOT be followed by other photos.

It is very important that you also remember what will be seen in this second phase. You will definitely need this information to finish the task successfully. This information will not be presented again, so remember it well.

Instructions Phase 2, counterconditioning condition. After the first phase, follows a second phase. During the second phase of the learning experiment, you will again see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of maimed bodies). Each photo will again be preceded by a photo that indicates which type of photo (positive or negative) will appear.

[bold] Watch out: There is an important difference between the first and the second phase:

If you see a photo of this product [display of Picture 2 or Picture 1, depending on counterbalancing condition] a positive photo will appear.

If you see a photo of this product [display of Picture 1 or Picture 2, depending on counterbalancing condition] a negative photo will appear.

It is very important that you also remember which product goes together with which type of photo (positive or negative) in this second phase. You will definitely need this information to finish the task successfully. This information will not be presented again, so remember well which product goes together with which type of photo in this second phase.

Rating instructions. Before we start with the learning experiment, you first have to indicate how pleasant you find the photos of the products, which will later appear. Make sure, however, that you don’t forget any of the instructions of the learning experiment that will follow!

Indicate for every photo of a product how positive (pleasant) or negative (unpleasant) your impression is. For every photo of a product, you have a scale ranging from -10 (very negative) to +10 (very positive). You can therefore make a very precise judgment. Click for every photo of a product on the value that fits best.

Please try your best to be as precise as possible. Earlier research has shown that this type of judgments can certainly lead to meaningful results.

Experiment 2a

Instructions Phase 1, all conditions. In a first phase of this learning experiment you will see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of maimed bodies). Each photo will be preceded by a meaningless word that indicates which type of photo (positive or negative) will appear.

If you see the word ENANWAL [UDIBNON], a positive photo will appear.

If you see the word UDIBNON [ENANWAL], a negative photo will appear.

It is very important that you remember which word goes together with which type of photo in this first phase.

You will definitely need this information to finish the task successfully. This information will not be presented again, so remember well which word goes together with which type of photo in this first phase.

Instructions Phase 2, extinction condition. After the first phase, follows a second phase.

During the second phase of the learning experiment, you will only see the words ENANWAL and UDIBNON without them being followed by photos.

It is very important that you also remember what will be seen in this second phase.

You will definitely need this information to finish the task successfully.

This information will not be presented again.

Instructions Phase 2, counterconditioning condition. After the first phase, follows a second phase.

During the second phase of the learning experiment, you will again see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of maimed bodies). Each photo will again be preceded by a photo that indicates which type of photo (positive or negative) will appear.

If in the second phase you see the word UDIBNON [ENANWAL], a positive photo will appear.

If in the second phase you see the word ENANWAL [UDIBNON], a negative photo will appear.

It is very important that you also remember which word goes together with which type of photo in this second phase.

You will definitely need this information to finish the task successfully. This information will not be presented again, so remember well which word goes together with which type of photo in this second phase.

Experiment 2b

Instructions Phase 1, all conditions. In a first phase of this learning experiment you will see pleasant, positive photos (e.g., of flowers) and unpleasant, negative photos (e.g., of maimed bodies). Each photo will be preceded by a meaningless word.

Please watch the photos and words attentively.

You don’t have to do anything else.

Instructions Phase 2, extinction condition. Now follows a second phase.

During the second phase of the learning experiment, you will only see the words ENANWAL and UDIBNON without them being followed by photos.

It is very important that you remember what will be seen in this second phase.

You will definitely need this information to finish the task successfully.

This information will not be presented again.

Instructions Phase 2, counterconditioning condition. Now follows a second phase.

During the second phase ENANWAL [UDIBNON] will be followed by negative photos and UDIBNON [ENANWAL] will be followed by positive photos.

It is very important that you remember what will be seen in this second phase.

You will definitely need this information to finish the task successfully

This information will not be presented again.

[1] Lipp et al. (2010) also asked participants to rate the extent to which a CS caused the presentation of the good or bad US. The mere instruction that CS-US pairings would be reversed (counterconditioning) or that the USs would no longer be presented (extinction) did influence these causal ratings but only slightly and to a much lesser extent than the actual experience of a change in contingencies. One could argue that if participants had fully processed and believed the instructions, there should have been a maximal change in causal ratings immediately after instructions. Hence, the fact that instructions had only a minimal effect on causal ratings can be seen as support for the idea that participants in the Lipp et al. studies did not process the instructions thoroughly.

[2] In this pilot study ($N = 20$) we used Pokemons as CS, which we later decided not to use because of a too strong evaluative connotation of the Pokemons independent from conditioning.

[3] Both for this and the following experiment, we also performed analyses based on only those participants who correctly indicated all pairings. In both studies, the pattern of results was similar to the pattern found with the whole sample. The most important difference was that the contrast between the extinction and the control condition in the reduced sample of Experiment 1 was not significant ($p = .139$). Please note that the power of this analysis is reduced due to the exclusion of participants, especially in the extinction condition. The sample of participants with incorrect memory was too small to allow for more systematic comparisons of participants with correct and incorrect memory.

[4] Participant numbers in the experiment with real pairings differ because it was originally planned to limit the analysis to participants with correct memory (see also Footnote 3). Therefore additional participants were tested in the extinction and counterconditioning conditions in order to compensate for exclusion of people with incorrect memory

[5] The evaluative ratings were always collected after the IAT and might thus be biased by forgetting, additional learning, or consolidation that occurs during the IAT (see Ebert, Steffens, von Stülpnagel, & Jelenec, 2009). Nevertheless, for exploratory reasons, we also analyzed the rating data. Most importantly, we found a significant main effect of valence, $F(1,164) = 141.04$, $p < .001$, $\eta^2_{partial} = 0.46$, indicating a preference in line with the instructions or actual pairings of the first phase. There was also an interaction of valence and instruction type, $F(2,164) = 5.35$, $p = .006$, $\eta^2_{partial} = 0.061$. Contrasts failed to reveal a significant difference between the EC effects in the control and the counterconditioning condition, $p = .283$. EC effects in the control and in the extinction condition differed, but in the direction opposite to what was expected, $p = .048$. In the ratings, the EC effect was larger in Experiment 2b (first phase experience-based) than in Experiment 2a (first phase instruction-based), $F(1,164) = 5.61$, $p = .019$, $\eta^2_{partial} = 0.033$.

[6] Please note that attributing an EC effect to propositional processes is not the same as claiming that a result from an EC procedure is due to demand compliance. Demand compliance requires not only that participants have conscious propositional knowledge of the CS-US relations but also that they intentionally use this knowledge in order to comply with perceived demands. Propositional knowledge could lead to changes in liking also in other ways (e.g., unintentionally or because participants use this knowledge to justify their preferences; see De Houwer et al., 2005; Gast et al., 2012). In the current research we used implicit measures in order to reduce the impact of demand compliance (see below).