A memory-based judgment account of expectancy-liking dissociations in evaluative conditioning

Frederik Aust[1], Julia M. Haaf[2], & Christoph Stahl[1]

[1] University of Cologne, Germany
[2] University of Missouri, MO

Postprint of manuscript accepted for publication at *Journal of Experimental Psycholog: Learning, Memory, and Cognition* on February 28th, 2018. 10.1037/xlm0000600

Abstract

Evaluative conditioning (EC) is a change in liking of neutral conditioned stimuli (CS) following pairings with positive or negative stimuli (unconditioned stimulus, US). A dissociation has been reported between US expectancy and CS evaluation in extinction learning: When CSs are presented alone subsequent to CS-US pairings, participants cease to expect USs but continue to exhibit EC effects. This dissociation is typically interpreted as demonstration that EC is resistant to extinction, and consequently, that EC is driven by a distinct learning process. We tested whether expectancy-liking dissociations are instead caused by different judgment strategies afforded by the dependent measures: CS evaluations are by default integrative judgments—summaries of large portions of the learning history—whereas US expectancy reflects momentary judgments that focus on recent events. In a counterconditioning and two extinction experiments, we eliminated the expectancy-liking dissociation by inducing nondefault momentary evaluative judgments, and demonstrated a reversed dissociation when we additionally induced nondefault integrative expectancy judgments. Our findings corroborated a-priori predictions derived from the formal memory model MINERVA 2. Hence, dissociations between US expectancy and CS evaluation are consistent with a single-process learning model; they reflect different summaries of the learning history.

Keywords: evaluative conditioning; extinction; counterconditioning; expectancy learning; evaluative judgment
Word count: 13,874

Evaluative conditioning (EC) is a change in liking of neutral *conditioned stimuli* (CS) following pairings with positive or negative *unconditioned stimuli* (US; De Houwer, 2011; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). For example, an initially neutral brand logo (the CS) that is repeatedly paired with positive stimuli (the USs) in advertisement settings is later evaluated more positively compared to initial evaluations or unpaired logos. In this sense, EC is considered to be a model of the effects of advertising (Biegler & Vargas, 2013), and of attitude acquisition in general (De Houwer, Thomas, & Baeyens, 2001).

Most human associative learning phenomena can be accounted for by a propositional process which presumably requires conscious awareness of the to-be-learned regularities—the CS-US contingencies—to affect behavior (Mitchell, De Houwer, & Lovibond, 2009). It has been argued, however, that EC violates this principle (Baeyens & De Houwer, 1995; De Houwer et al., 2001). Multiple studies report that EC may occur without conscious awareness of CS-US contingencies (Lovibond & Shanks, 2002; see Sweldens, Corneille, & Yzerbyt, 2014 for a recent review). Moreover, EC has been claimed to be resistant to extinction and, hence, to occur despite conscious awareness of the absence of CS-US contingencies (Baeyens, Crombez, Van den Bergh, & Eelen, 1988; Baeyens, Dı'az, & Ruiz, 2005; Dwyer, Jarratt, & Dick, 2007; Hermans, Crombez, Vansteenwegen, Baeyens, & Eelen, 2002; Vansteenwegen, Francken, Vervliet, De Clercq, & Eelen, 2006). This dissociation between expectancy and liking cannot be readily explained by the aforementioned propositional learning process; hence, EC was taken to involve distinct processes that differ from those underlying other associative learning phenomena (De Houwer et al., 2001). Consequently, dual-process theories of attitude acquisition, which postulate an additional, automatic, associative learning process (e.g., Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000), have become popular among EC theorists.

Recent research has, however, cast doubt on whether these critical findings hold, and if they do, whether a dual-process account is in fact necessary to explain them. Overall, the evidence for EC without CS-US contingency awareness is weak, with unintentional (incidental) EC perhaps best supported by the data (Corneille & Stahl, n.d.; Heycke, Aust, & Stahl, 2017; Heycke, Gehrmann, Haaf, & Stahl, 2018; Stahl et al., 2016; Sweldens et al., 2014). The present study investigates a recent single-process account of why EC appears to be resistant to extinction (Lipp, Mallan, Libera, & Tan, 2010; Lipp, Oughton, & LeLievre, 2003; Lipp & Purkis, 2006). We briefly review the central findings obtained with extinction procedures in EC research and then suggest a parsimonious single-process model for these findings. Next, we present three experiments that tested the model predictions in a counterconditioning and an extinction procedure.

Resistance to extinction

Until recently, the majority of EC studies supported the interpretation that EC is resistant to extinction (e.g. Baeyens et al., 1988, 2005; Hermans et al., 2002). For example, Hermans et al. (2002) report a dissociation between CS evaluation and US expectancy in two experiments. The authors used a common extinction procedure, in which CSs were

---

Frederik Aust and Christoph Stahl, Department Psychology, University of Cologne, Cologne, Germany; Julia M. Haaf, Department of Psychological Sciences, University of Missouri, Columbia, MO, USA. The data reported here were previously presented at 59th Conference of Experimental Psychologists, Dresden, Germany, and 5th European Meeting on the Psychology of Attitudes, Cologne, Germany. FA, JH, and CS planned research; FA analyzed the data and performed the simulations; FA, JH, and CS wrote the manuscript. We thank Hannah Frings, Julia Krasko, Katharina Mattonet, Philipp Musfeld, Marius Schmitz, and Lisa Spitzer for their help with data collection.

Correspondence concerning this article should be addressed to Frederik Aust, Department Psychology, Herbert-Lewin-Str. 2, 50931 Cologne, Germany. E-mail: frederik.aust@uni-koeln.de

paired with USs in an acquisition phase and presented alone in a subsequent extinction phase. To assess the effect of the extinction procedure on EC, they compared CS evaluations obtained after acquisition to those obtained after extinction. Hermans et al. (2002) found that EC was unaffected by the extinction phase, whereas US expectancy was extinguished.

The resistance of CS evaluation to extinction stands in contrast to the rapid extinction of conditioned responses observed in Pavlovian conditioning (Lovibond, 2004) and human associative learning more generally (Mitchell et al., 2009). Dissociations between US expectancy and CS evaluation, as those reported by Hermans et al. (2002), pointedly illustrate this contrast and are central to the debate between single- and dual-process learning theorists (e.g., Baeyens, Vansteenwegen, & Hermans, 2009). The latter suggests that, unlike US expectancy, CS evaluation is driven by a distinct learning process that presumably reflects temporo-spatial co-occurrences between CSs and USs (CS-US contiguity) but not the predictive value of a CS (statistical CS-US contingencies; e.g., Sweldens et al., 2014). Hence, the dual-process account posits that a change in the predictive value of a CS, for example due to extinction learning, affects US expectancy but not CS evaluation.

In 2010, a meta-analysis by Hofmann et al. (2010) rekindled this debate: their results indicated that EC is not, strictly speaking, resistant to extinction. They found a substantial reduction of EC in studies that assessed the EC effect both after acquisition and then again after extinction. CS-alone trials reduced the EC effect by 37% (i.e., from $d = 0.85$ to $d = 0.53$). This finding is difficult to reconcile with conventional dual-process theories of EC and, thus, motivated new research on the topic.

Gawronski, Gast, and De Houwer (2014), for example, suspected that extinction of EC may be dependent on characteristics of the study procedure and tested their hypothesis experimentally. They found no extinction when they compared the EC effect between different groups of participants who evaluated CSs only once, either after the acquisition or after the extinction procedure. EC was (partly) extinguished only when participants evaluated CSs twice—after the acquisition *and* after the extinction phase. Extinction was, however, only observed in explicit evaluative ratings but not in affective priming. Based on their results, Gawronski et al. (2014) argued that these changes in CS evaluation do not reflect genuine changes in liking. Instead, they argued that specific judgment-related nuisance processes (e.g., due to repeated CS evaluation) may be responsible for the artifactual extinction of EC in explicit evaluative ratings; the true underlying evaluative representation was assumed to be unaffected by the extinction procedure, as supported by the presumably less obtrusive evaluative priming measure. Gawronski et al. (2014) argued that their finding resolves the contradiction between the extinction effect found by Hofmann et al. (2010) and the resistance to extinction predicted by dual-process theories. They concluded that any theoretical account has to explain how EC is largely resistant to extinction.

Extinction learning is a prominent example of such an expectancy-liking dissociation. A similar dissociation has been reported in counterconditioning procedures, in which CSs are associated with USs of opposing valence in two subsequent parts of the learning procedure (Lipp et al., 2010; Lipp & Purkis, 2006): At the end of this two-part learning procedure, participants exhibited no EC effects, although they continued to expect CSs and USs to co-occur according to the regularity learned in the more recent second part. This dissociative

pattern is the opposite of the one observed in extinction procedures, at the end of which participants (continue to) exhibit EC effects but no longer expect the CS to co-occur with USs. Dual-process theories explain this dissociation just as they explain the dissociation in extinction procedures: The associative learning process assumedly is driven by CS-US contiguity and, thus, the predictive value of CSs is irrelevant to their evaluation (Sweldens et al., 2014). Hence, no EC is to be expected because CSs are paired with positive as often as with negative USs.

The temporal integration hypothesis

Single-process theories cannot explain resistance to extinction, and particularly the dissociation between US expectancy and CS evaluation, by referring to distinctive properties of separate learning systems. Additional assumptions are necessary. Lipp et al. (2010) discussed a set of auxiliary assumptions for the single-process account, which we will refer to as the *temporal integration hypothesis*, to account for the expectancy-liking dissociation (also see Lipp et al., 2003). They argue that US expectancy and CS evaluations reflect different summaries of the same underlying representation. The assumption is that memory stores a unitary representation of the CS-US pairing, and that the learning history is conserved and organized along a temporal dimension or by contextual properties. Moreover, it is assumed that memory for CS-US pairings can be flexibly used to meet the (assumed) task demands. Lipp et al. (2010) argue that, by default, CSs are evaluated under consideration of the entire learning history—participants make *integrative* evaluative judgments. In contrast, predictions or judgments of US expectancy are made by default in reference to recent events—participants make *momentary* expectancy judgments. These opposite default judgment strategies are assumed to be afforded by the tasks. Thus, Lipp et al. (2010) proposed that the expectancy-liking dissociation is not indicative of two independent learning systems; instead, they propose that the dissociation is caused by different default judgment strategies underlying US expectancy versus CS evaluation responses.

The temporal integration hypothesis is inspired by a very similar idea proposed in the field of causal learning. Collins and Shanks (2002) found a dissociation between causal strength judgments and outcome prediction. Participants viewed a series of trials showing imaginary laboratory records that documented butterfly species' reactions to radiation exposure. Radiation caused genetic mutations in half of the butterflies and prevented mutations in the others. Akin to a counterconditioning procedure, these contingencies reversed in the middle of the experiment. Thus, across all trials there was no causal relationship between radiation and mutation for any butterfly species. Similar to US expectancy ratings in EC experiments, intermittent predictions about the occurrence of genetic mutations closely mirrored the changes in contingencies: Participants correctly predicted that radiation would first cause but later prevent mutations, or vice versa. But end-of-study causal strength ratings dissociated from participants' last predictions: In causal strength ratings participants favored neither cause nor prevention. Collins and Shanks (2002) further found that this dissociation was affected by the frequency of these ratings. When participants repeatedly rated causal strength throughout the learning procedure, their end-of-study ratings corresponded to their intermittent predictions.

To explain their findings, Collins and Shanks (2002) argued that participants can flexibly adopt different judgment strategies. For frequent judgments, a momentary strategy is adopted in which ratings reflect only the most recent information (i.e., that has been acquired since the last judgment). In contrast, when judgments are made only at the end of a series of events, an integrative strategy is adopted, in which ratings incorporate information from the entire event series. Rather than being dichotomous, these strategies can be thought of as smaller or larger averaging windows used to aggregate information across time. Matute, Vegas, and De Marez (2002) explored factors that cause participants to adopt a momentary or integrative judgment strategy. They found that questions targeting the predictive value of a stimulus induced momentary judgments, whereas questions about contiguity and causality induced integrative judgments. Moreover, they were able to manipulate the adopted judgment strategy via postexperimental instructions. In short, this research implies that participants flexibly use the learned information to meet the (assumed) demands of the task set by the experimenter.

Building on the research by Collins and Shanks (2002), Lipp and Purkis (2006) found that dissociations between US expectancy and CS evaluations are similarly affected by the frequency of evaluative ratings. In a counterconditioning and an extinction procedure, participants provided pleasantness ratings either twice (i.e., after each of the two learning phases) or only once at the end. When only one final rating was collected, participants' ratings reflected averages across the entire learning procedure: In the counterconditioning procedure, participants exhibited no EC effect; whereas in the extinction procedure, they exhibited a robust EC effect. In contrast, when participants provided multiple ratings, their CS evaluations reflected only the most recent CS-US contingencies, and the expectancy-liking dissociation was eliminated: In the counterconditioning procedure, participants reported causal relationships between CSs and USs in accord with the contingencies inherent in the respective part of the procedure, and they exhibited EC effects corresponding to these causal judgments. In the extinction procedure, participants reported no causal relationship after the extinction phase, and, correspondingly, they now also failed to exhibited an EC effect. In other words, the final expectations no longer dissociated from end-of-study evaluations—the EC effect was successfully extinguished. Notably, the extinguished EC effect reappeared when participants were asked to evaluate the CSs again in a different context and response format at the end of the study. Lipp et al. (2010) argued that their findings can be explained by the temporal integration hypothesis. They proposed that, when asked repeatedly throughout the learning procedure, participants made momentary judgments that reflected recent trends in CS-US contingencies (i.e., showed EC in counterconditioning but not after extinction). On the other hand, postexperimental pleasantness ratings in a different context and response format were by default integrative judgments that reflected the entire learning history (i.e, showed no EC in counterconditioning but did show EC after extinction).

The temporal integration hypothesis may reconcile expectancy-liking dissociations with single-process theories of EC, but the proposed auxiliary assumptions need to be tested rigorously. Previous research leaves room for alternative explanations of the extinction of EC, and it has not tested the effects of judgment strategies of US expectancy and CS evaluation concurrently. Here we address those shortcomings. We tested two predictions from the hypothesis' core assumptions more stringently and without allowing for alternative accounts

in terms of demand effects induced by multiple pleasantness judgments (as proposed, e.g., by Gawronski et al., 2014). Remember that Lipp and Purkis (2006) elicited nondefault momentary CS pleasantness judgments by collecting ratings intermittently during the learning procedure. As argued by Gawronski et al., multiple intermittent CS evaluations could alter the evaluative learning process or bias response behavior by inducing demand characteristics and thereby artificially create momentary judgments. The present studies avoided this potential confound by collecting CS evaluations (as well as, in Experiment 3, expectancy judgments) only after the learning phase and thereby eliminate alternative explanations in terms of demand characteristics.

If previous findings are indeed caused by judgment strategies, then (1) it should be possible to manipulate these strategies for US expectancy and CS evaluation after the learning procedure and without intermittent CS pleasantness judgments. Moreover, if the default judgment strategies for CS pleasantness (Lipp & Purkis, 2006) and US expectancy (Collins & Shanks, 2002; Matute et al., 2002) are malleable, (2) the expectancy-liking dissociation in extinction learning should be reversible if one could elicit the opposite nondefault judgment strategies. A concurrent cross-over manipulation of judgment strategies for both US expectancy and CS pleasantness would predict a double-dissociation pattern. Combined in a single experiment, this constitutes a rigorous test of the temporal integration hypothesis. Confirmation of these double-dissociation predictions, while eliminating alternative accounts, would provide stronger support for the temporal integration hypothesis that goes well beyond that provided by previous studies.

MINERVA 2: A candidate single-process model

The temporal integration hypothesis does not specify how the learning history is conserved, how temporal organization is achieved, or how the information is summarized to perform judgment tasks. Mitchell et al. (2009) postulated that human associative learning is based on memory for past events. They further suggested that MINERVA 2 (Hintzman, 1984, 1986, 1988), a simple but popular model of episodic memory, may be the simplest model consistent with a memory system supporting their propositional single-process view of human associative learning (p. 187, Mitchell et al., 2009) (see also De Houwer, 1998; Klauer, 2009). In an attempt to fill in the blanks of the temporal integration hypothesis, we followed the suggestion by Mitchell et al. (2009) and adopted the memory architecture formalized in MINERVA 2. We explore the theoretical position that US expectancy and CS evaluation are memory-based judgments that rely on a unitary representation of CS-US pairing episodes. Using a formalized model enables us to make more specific predictions than current process theories of EC.

MINERVA 2 assumes that each CS-US pairing is stored as a trace in a unitary memory system. Episodes are encoded in a feature-based manner. Each memory trace consists of a series of slots, each of which indicates whether a feature is present (or absent) in a given episode. In the present application, subsets of these feature slots are dedicated to CS, US, and context features. When memory is probed (i.e., when a judgment is made), the stimulus and context features of the probe are simultaneously compared to all traces in memory. Each memory trace is activated according to its similarity to the memory probe. The recalled

memory content is computed as a weighted average of all memory traces, where similar and strongly activated traces receive a larger weight than dissimilar and weakly activated traces. Hence, the recalled information is a mixture of all memory traces—rather than reflecting one specific past episode.

In line with current theorizing in memory research (e.g., Howard & Kahana, 2002; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), we assume that the unitary memory system holds information about the (temporal) context of all stored episodes. MINERVA 2 is not equipped with a dedicated mechanism to impose a temporal structure on the stored episodes but such an organization can be achieved by assuming that a changing context is encoded in each episode. This conceptualization is consistent with mechanisms proposed in perceptual and memory research, where it is suggested that the continuous flow of information is automatically segmented and structured into discrete events (Zacks et al., 2007). Matute, Lipp, Vadillo, and Humphreys (2011) have similarly invoked the concept of temporal contexts in research on associative learning. They found that participants spontaneously (i.e., without instructions) structure learning procedures by creating temporal contexts. Participants then used these contexts to retrieve associative information to guide their behavior and inform their prediction of future events. Thus, we assumed that the temporal organization of the learning history is retained via (perceived or internally generated) contexts that structure the incoming information into meaningful events. These assumptions allowed us to derive specific predictions for the learning procedures implemented in the present studies.

The present study

The overarching goal of this research was to test whether a single-process learning account can explain the expectancy-liking dissociation in EC. Building on the work by Collins and Shanks (2002) as well as Lipp and Purkis (2006), we tested the temporal integration hypothesis (Lipp et al., 2010, 2003), which posits that US expectancy and CS pleasantness judgments are different summaries of a common underlying representation of CS-US pairings. We attempted to modify the default momentary and integrative judgment strategies for US expectancy and CS pleasantness judgments after completion of the learning procedure and without intermittent judgments. Moreover, we aimed at *reversing* the expectancy-liking dissociation in extinction learning by inducing nondefault integrative US expectancy and momentary CS pleasantness judgments.

We conducted one counterconditioning and two extinction experiments (see Table 1 for an overview).[1] For the counterconditioning procedure in Experiment 1, CS-US pairings were presented in two contexts: CSs were paired with positive USs in the first, and with negative USs in the second context, or vice versa. During the learning procedure, participants provided intermittent US expectancy ratings. After learning, participants judged CS pleasantness either without reference to learning contexts (to elicit default integrative judgments) or

[1]We ran three additional experiments as part of this project, which will be reported elsewhere. The data are available at https://github.com/methexp/rawdata.

for a specific context (to elicit momentary judgments). Experiment 2 used an extinction procedure and presented half of the CSs together with USs in the first but alone in the second context. To hold the number of USs constant across contexts, the other half of the CSs was presented alone in the first but with USs in the second context, thereby implementing a concurrent acquisition procedure. Experiment 3 replicated and extended Experiment 2: Participants provided no intermittent judgments but rated US expectancy only after completion of the learning procedure, either for both learning contexts together (to elicit integrative judgments) or for a specific context (to elicit default momentary judgments).

.

Experiment 1

In parallel with Collins and Shanks (2002) and Lipp and Purkis (2006), we first tested the temporal integration hypothesis and our memory-based judgment simulation of EC with the expectancy-liking dissociation in a counterconditioning procedure. If the assumptions of temporal integration hypothesis hold, single-process theories of EC can account for this dissociation by assuming that end-of-study CS evaluations are integrative judgments, and accordingly, no EC effect is to be expected because the effects of positive and negative CS-US pairings cancel each other out.

We first designed a simulation of a simplified counterconditioning procedure to generate more specific predictions using MINERVA 2 (for details see Appendix A). One CS was first paired with a positive and then with a negative US, conversely, a second CS was first paired with a negative and then with a positive US; a third CS was paired with a neutral US. Moreover, we simulated context changes in between the first and second phase of the learning procedure as well as prior to end-of-study CS pleasantness ratings. Thus, we assumed participants would experience the end-of-study rating procedure as different from the learning procedure. To predict US expectancy and CS pleasantness ratings, we reasoned that the CS in question and the current context act as cues to recall previous pairings with USs. If the recalled memory content was positive we predicted an expectation of a positive USs and a positive CS evaluation. We, thus, predicted US expectancy and CS pleasantness ratings based on the same information.

Our simulation predicted a pattern of results consistent with the temporal integration hypothesis, Figure 1A. During the learning procedure, the valence of the recalled memory content closely followed the CS-US contingencies. The recalled memory contents acquired the USs' valence but due to the context change the CS-US pairings in the counterconditioning phase quickly reversed the contents' valence. Thus, for the last trial the simulation predicted expectation of the US that had been paired with a given CS in the second context.

More importantly, the same pattern was predicted for end-of-study judgments when the learning contexts were reinstated. For example, when a CS that had first been paired with a positive and then with a negative US was presented in the first context, the valence of the retrieved memory contents was positive. However, when the same CS was presented in the second context, the recalled information was negative. The reinstated context features promoted the activation of memory traces of episodes from the respective context. This

contextualized retrieval of CS-US pairings assumedly underlies momentary judgments of US expectancy and CS pleasantness. For the new context—when no learning context was reinstated—our simulation predicted that episodes from both contexts contributed equally to the retrieved memory contents. Positive and negative CS-US pairings effectively cancelled each other out. Thus, the simulation predicted no EC effect in default integrative end-of-study pleasantness ratings.

To conclude, in line with the temporal integration hypothesis, the simulation of the counterconditioning procedure predicted momentary judgments in intermittent US expectancy ratings and both momentary and integrative judgments in end-of-study CS pleasantness ratings, depending on context cues. Hence, our single-process memory model simulation produced an expectancy-liking dissociation, which has been taken as evidence for dual-process theories of EC: Marked US expectancies in the last trial but no EC effect in end-of-study CS pleasantness ratings for the new context. It, nonetheless, predicted EC effects in momentary end-of-study CS pleasantness ratings when learning contexts are reinstated. Therefore, no expectancy-liking dissociation is expected when comparing momentary US expectancy to momentary CS pleasantness ratings.

We designed an experiment to test these predictions. We conducted a counterconditioning experiment with intermittent US expectancy and end-of-study CS pleasantness ratings in different contexts. We showed participants a stream of pictures in which CSs were first paired with positive and later with negative USs, or vice versa. In contrast to Lipp and Purkis (2006) we asked participants to evaluate CSs only after completion of (rather than repeatedly during) the learning procedure. This procedural change ruled out that intermittent CS pleasantness judgments affected the evaluative learning process and artificially induced subsequent momentary judgments (e.g., via conversational logic demands). Participants provided end-of-study CS pleasantness ratings without reference to learning contexts (to elicit default integrative judgments) and for each of the learning contexts (to elicit nondefault momentary judgments). We expected (1) to observe the predicted expectancy-liking dissociation between intermittent US expectancy ratings in the last trial on the one hand and integrative end-of-study CS pleasantness ratings on the other hand, but (2) to eliminate the expectancy-liking dissociation by demonstrating EC effects that mirror US expectancy ratings in momentary end-of-study CS pleasantness ratings.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. The simulation code, experimental software[2] and materials, data, and analysis scripts[3] are available at https://osf.io/vnmby/.

---

[2]We created all experiments in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012).

[3]We used R (Version 3.5.1; R Core Team, 2017) and the R-packages *afex* (Version 0.21.2; Singmann, Bolker, Westfall, & Aust, 2017), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2015), *emmeans* (Version 1.3.0; Lenth, 2018), and *papaja* (Version 0.1.0.9842; Aust & Barth, 2017) for all analyses and reporting.

Participants. We recruited 40 participants from our lab participant database via e-mail for this experiment. Eligible volunteers were 18-60 years old, fluent in German and (according to our database) had not participated in any studies on evaluative conditioning for at least one year. Participants who aborted the experiment were not included in the analyses. The sample size was determined informally based on previous experience with EC experiments. The data of three participants were lost due to a technical error leaving the data of 37 participants for analysis. Participants' mean age was 23.69 years ($SD = 6.50$), 26 were female, 11 studied psychology or media psychology, all participants declared intact color vision, and 9 reported to have had prior knowledge about the CS pictures. We compensated participants with 8€ or course credit.

Apparatus and material. We conducted the experiment in five dimly lit and sound-attenuated booths and presented all stimuli on a 17" CRT-monitor.

Because a seemingly random stimulus sequence, a large proportion of filler stimuli, and a low proportion of valent stimuli, have been reported to be conducive to (associative) EC (Jones, Fazio, & Olson, 2009), our learning procedure consisted of a mixture of critical CS-US pairings and irrelevant filler trials. Critical CS-US pairs consisted of 12 neutral cartoon characters taken from Stahl and Heycke (2016) as CSs and 12 positive or 12 negative low-arousal IAPS pictures as USs (Lang, Bradley, & Cuthbert, 2008, Table C1). All positive USs were pictures of animals; all negative USs were pictures of humans. We introduced the confound between US valence and category because we wanted to rule out that intermittent US expectancy judgments affected the evaluative learning process—the confound enabled us to assess US expectancy without referring to US valence.

The filler trials consisted of six neutral CS-US pairs, three CS-CS pairs, three individual CSs, three US-US pairs, as well as three individual USs of intermixed valence, and six blank screens. For filler CSs, we used additional cartoon characters (Stahl & Heycke, 2016) and for filler USs, we used IAPS pictures from two additional US categories. All neutral USs depicted household items, and the intermixed USs depicted natural scenes (see Table C1). The filler stimuli were included to make contingency learning more demanding, and to obscure the confound between US valence and categories and thereby to further mitigate possible effects of intermittent ratings on evaluative learning.

US expectancy has previously been assessed with predictive ratings (e.g. of the extent to which participants expected a US following the presentation of a CS, p. 224 Hermans et al., 2002; also see Vansteenwegen et al., 2006) or causal questions ("To which extent (0–100%) does [the CS] cause the [US] to appear?", Lipp et al., 2010)(also see Collins & Shanks, 2002; Lipp & Purkis, 2006). In the context of contingency learning, Matute et al. (2002) found that predictive and causal questions elicit comparable integrative judgments but that predictive questions more effectively elicit momentary judgments. Hence, we employed a predictive question: "Next time this creature is presented, what type of picture will it be shown with?" Participants provided probability estimates for animal, human, and object on an eleven-point scale ranging from 0% to 100%.

We collected CS pleasantness ratings on an 19-point scale ranging form *very unpleasant* to *very pleasant*. To assess memory for CS-US pairs, we separately tested recognition memory

for US category and US identity for each CS. For US category recognition, we presented individual CSs and participants selected a US category in an 3-alternative forced-choice (3-AFC) task (e.g., "animal", "human", or "object"). For US identity recognition, participants selected one US out of all USs from the correct US category in an 12-AFC. We also performed a funnel debriefing to assess the extent to which participants were aware of the purpose of the study and the hypotheses. The debriefing served to inform the design of future incidental-learning studies; they are irrelevant to the present hypotheses.

Procedure and design. After obtaining informed consent, participants filled in demographic information about gender, age, handedness, field of study, and visual impairments. We then instructed participants that we would present a stream of pictures in 2×3 blocks and asked them to attend the stream carefully, to detect regularities, and to memorize repeating pairs of pictures (for similar instructions see e.g., Kattner & Green, 2015; Moran & Bar-Anan, 2013; Richter & Gast, 2017; Zanon, De Houwer, & Gast, 2012). We warned that, during the course of the study, we would test whether they had continuously attended the stream. To distract from the contingency between CS and US valence, we pretended that we were interested in participants' vigilance while they monitored images from surveillance cameras.

Table 1

Illustration of our learning procedures in Experiments 1-3

	Learning procedure		Intermittent ratings	End-of-study ratings	
	First context ($\mathcal{A}$)	Second context ($\mathcal{B}$)	US expectancy	CS pleasantness	US expectancy
Experiment 1					
Counterconditioning	CS_1 US_+	CS_1 US_-	CS_1 $(\mathcal{A},\mathcal{B})$	CS_1 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
	CS_2 US_-	CS_2 US_+	CS_2 $(\mathcal{A},\mathcal{B})$	CS_2 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
Experiment 2					
Acquisition	CS_1	CS_1 US_+	CS_1 $(\mathcal{A},\mathcal{B})$	CS_1 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
	CS_2	CS_2 US_-	CS_2 $(\mathcal{A},\mathcal{B})$	CS_2 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
Extinction	CS_3 US_+	CS_3	CS_3 $(\mathcal{A},\mathcal{B})$	CS_3 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
	CS_4 US_-	CS_4	CS_4 $(\mathcal{A},\mathcal{B})$	CS_4 $(\mathcal{C},\mathcal{B},\mathcal{A})$	–
Experiment 3					
Acquisition	CS_1	CS_1 US_+	–	CS_1 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$	CS_1 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$
	CS_2	CS_2 US_-	–	CS_2 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$	CS_2 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$
Extinction	CS_3 US_+	CS_3	–	CS_3 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$	CS_3 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$
	CS_4 US_-	CS_4	–	CS_4 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$	CS_4 $(\mathcal{A} \parallel \mathcal{B} \parallel \mathcal{C})$

Note. Calligraphic font denotes context features; mapping of features to context was counterbalanced in all experiments. In Experiments 1 and 2, participants provided end-of-study ratings for every context (within-subject manipulation) but for only one randomly selected context in Experiment 3 (between-subject manipulation). CS = Conditioned stimulus; US = Unconditioned stimulus

The conditioning procedure consisted of two phases, Table 1. In the initial acquisition phase, we paired 6 critical CSs with positive and the remaining 6 critical CSs with negative USs. In the subsequent counterconditioning phase, critical CSs were paired with USs of the opposite valence. CSs were randomly assigned to one of the two US valence orders. Filler CSs that were paired with neutral USs in the first phase were paired with new neutral USs in the second phase.

We created different contexts for the first and second phase to standardize participants' temporal organization of the learning history (Matute et al., 2011) and facilitate later reference to each phase in targeted questions about particular portions of the learning procedure. The context features—background color and CS position—were randomly assigned to the first or second phase for each participant. The background color of the screen was either yellow or blue; CSs were presented either on the left or right side of the screen, with USs on the opposite side.

Both phases consisted of three subblocks, interrupted by self-paced breaks. In each of the subblocks, we presented all critical and filler trials three times. The stimulus sequence entailed no immediate stimulus repetitions but was otherwise random. In each trial, CSs were presented alone for 500 ms and then jointly with USs for another 1000 ms. Each CS was paired with only one US to facilitate accurate memory for pairings. For CS-CS or US-US pairs, one stimulus was randomly chosen to act as CSs; the second stimulus acted as US. There was no delay between trials (Jones et al., 2009). The conditioning procedure consisted of 648 trials (216 with critical CSs), and lasted approximately 20 minutes.

During the learning procedure, we intermittently presented CSs and asked participants to report their current US expectancy: "With what probability would you expect a photograph of a human [animal/object] with this creature?" In each subblock, we randomly selected six of the 18 CS-US pairs (including neutral pairs). Participants made US expectancy judgments on a random trial following the third and final presentation of the selected CS-US pair in each subblock. (i.e., ratings in the first subblock reflected participants' expectations after three CS-US pairings, ratings in the second subblock reflected participants' expectations after six, etc.). In the subblocks of the subsequent counterconditioning phase, we used the same CS-US pair selection as in the acquisition phase. For example, if we selected a CS-US pair for US expectancy ratings in the first subblock of the acquisition phase, we selected the same pair in the first subblock of the counterconditioning phase. Thus, participants reported their US expectancy twice for every CS-US pair, and three subblocks elapsed before the second rating. Each participant provided 36 US expectancy ratings, yielding 3 ratings per experimental condition.

Following the learning procedure, participants provided pleasantness ratings for each CS. Akin to the postexperimental rating condition by Lipp and Purkis (2006), we collected a first rating in a new context. In this new context, we presented CSs in the center of the screen on a black background and asked "How pleasant or unpleasant do you find this creature[, currently]?" That is, there was no reference to learning contexts. We then collected CS pleasantness ratings for the context of the second and then of the first phase. We reinstated the respective context features (background color and CS position) and asked participants "How pleasant did you find this creature during the second [first] half?" Each participant

provided 54 pleasantness ratings (including CSs from neutral CS-US pairs), yielding 3 ratings per experimental condition.

Next, we assessed participants' memory for CS-US pairs. We tested pairing memory for the second and then for the first phase. The order served to minimize memory interference and because pairing memory for the counterconditioning phase was of particular interest. After every response, we immediately tested participants' US identity recognition for the same CS-US pair. We probed memory for CS-US pairs in a new random order for each participant and context. Each participant provided 36 US category and US identity recognition responses (including neutral CS-US pairs), yielding 6 responses per experimental condition.

Finally, we administered the funnel debriefing, participants rated the pleasantness of each US category (human, animal, and object) from memory (i.e., the USs were not presented again), and indicated whether they had previously been familiar with the cartoon characters. On average, participants took 56.53 minutes ($SD = 18.02$) to complete the study.

Due to an error in the randomization procedure, we used the same assignment of CSs to US valence orders for all participants, with two consequences. First, the CSs assigned to the US valence orders systematically differed in pleasantness: CSs that were first paired with negative and later with positive USs were more pleasant a priori ($M = 5.52, SD = 0.92$) than CSs first paired with positive and later with negative USs ($M = 4.38, SD = 1.00$). This confound is unlikely to endanger our conclusions because it works against our predictions of an EC effect in the acquisition context and the absence of an EC effect in the new context. Our remaining predictions largely concern changes in CS pleasantness across contexts within each set of CS-US pairs, for which the confound is irrelevant. Second, CSs were paired with a random US in the acquisition phase, but in the counterconditioning phase some specific CS-US pairs were more likely than others. Mean US pleasantness and arousal were however comparable across conditions and closely matched the means of all USs of the corresponding category (Table C2). In sum, the error in the randomization procedure is vexing and subpar but unlikely to affect our results or conclusions.

Data analysis

For all analyses, we averaged participants' responses across items. We combined US expectancy ratings for each of the three US categories into a single measure of expectancy of the correct US by subtracting the ratings for incorrect categories from those for the correct category. For example, for CSs paired with pictures of objects, we calculated a US expectancy score $\bar{x}_{\text{expectancy}} = \bar{x}_{\text{object}} - (\bar{x}_{\text{human}} + \bar{x}_{\text{animal}})$ for every participant in every cell of the experimental design.

We performed ANOVAs and base our inference on p values and 95% confidence intervals as well as Bayes factors. For the frequentist analyses we always report Greenhouse-Geisser corrected degrees of freedom. For planned contrasts and post hoc comparisons, we compared least squares means (Lenth, 2018). To infer equivalence between two condition means, we performed two one-sided t tests (TOST; Lakens, 2017; Rogers, Howard, & Vessey, 1993; Wellek, 2002). In the TOST procedure the analyst defines a region of equivalence around the null value. She compares the mean difference to the upper and lower bound of this

region of equivalence using one-sided tests or a 90% confidence interval. The means are deemed equivalent if the difference between them is significantly larger than the lower bound and significantly smaller than the upper bound. For reasons of brevity only the result of the test that yields the larger p value is reported. Thus, in case of a significant TOST the analyst rejects the hypothesis that an effect is of a given size or larger. We adopted symmetric equivalence regions in units of standardized mean differences for within-participant comparisons of $\Delta \pm 0.3 d_r$ to reject small effects (Lakens, 2017). The α-level for all frequentist analyses was .05; p values were corrected for multiple comparisons where applicable.

For Bayesian ANOVAs we used default multivariate Cauchy priors with a scaling parameter of $r = 0.5$ on the fixed effects (Rouder, Morey, Speckman, & Province, 2012); for Bayesian t tests we used a default Cauchy prior with a scaling parameter of $r = \sqrt{2}/2$ on the effect size d_z (Rouder, Speckman, Sun, Morey, & Iverson, 2009). All Bayes factors were estimated to a precision of ±5%. Bayes factors quantify the evidence for an effect relative to the null hypothesis of no effect in the data at hand (e.g, Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). We use BF_{10} to denote evidence for an effect relative to the null hypothesis of no effect and BF_{01} to denote evidence for the null hypothesis of no effect relative to an effect. For example, $\mathrm{BF}_{10} > 1$ is evidence for the presence of an effect, whereas $\mathrm{BF}_{01} > 1$ is evidence for the absence of an effect. Bayes factors are readily interpretable as a graded measure of evidence. We will, however, follow the suggestion by Kass and Raftery (1995) to consider $1/3 < \mathrm{BF} < 3$ "not worth more than a bare mention" (p. 777). We do not report our prior beliefs in the hypotheses described here (prior odds; for discussion see Rouder et al., 2012). The interested reader may form their own prior beliefs and use the reported Bayes factors to determine their posterior belief in the hypotheses.

Results

In the following, we focus on the results for US expectancy and CS pleasantness ratings. See Appendix B for analyses of participants' CS-US pairing memory.

US expectancy. We analyzed expectancies of the correct US using a 2 (*US valence order*: US+ US− vs. US− US+) × 2 (*Context*: First vs. Second) × 3 (*Pairings*: 3 vs. 6. vs. 9) repeated-measures ANOVA. To facilitate the comparisons between predicted and observed US expectancy as well as between US expectancy and CS pleasantness, Figure 1B depicts a difference score between expectancies of positive and negative US.

As expected, participants quickly learned the CS-US contingencies. The number of repetitions of CS-US pairings affected expectancy of the correct US, $F(1.59, 57.2) = 25.45$, $MSE = 0.22$, $p < .001$, $\hat{\eta}^2_G = .073$, $\mathrm{BF}_{10} = 2.31 \times 10^8$. Follow-up tests indicated that expectancy of the correct US increased from three to six CS-US pairings, $\Delta M = .27$, 95% CI $[.19, \infty]$, $t(72) = 5.56$, $p < .001$, $\mathrm{BF}_{10} = 4.14 \times 10^5$ (one-tailed), but remained unchanged from six to nine CS-US pairings, $\Delta M = .05$, 90% CI $[-.04, .15]$, $t(72) = -2.07$, $p = .042$ (equivalence test adjusted for two comparisons), $\mathrm{BF}_{01} = 2.49$. There was weak evidence that our experimental manipulations had no other effects, all $p \geq .145$, all $\mathrm{BF}_{01} \geq 3.89$. To conclude, participants' expectancy for the correct US built up during and reached a plateau

toward the end of each learning phase. At the end of the experiment participants expected CSs to be accompanied by the US that they had last been paired with.

CS pleasantness. We analyzed CS pleasantness ratings using a 2 (*US valence order*: US+ US− vs. US− US+) × 3 (*Referenced context*: First vs. Second vs. None) repeated-measures ANOVA.

As predicted, referring to and reinstating learning contexts affected CS pleasantness ratings differently depending on US valence order, $F(1.18, 42.31) = 17.63$, $MSE = 10.32$, $p < .001$, $\hat{\eta}^2_G = .083$, $\text{BF}_{10} = 2.19 \times 10^6$, Figure 1B. Follow-up tests provided some evidence that in the new context participants made comparable CS pleasantness ratings for both US valence orders, $\Delta M = -0.12$, 90% CI [−1.15, 0.91], $t(103.46) = -1.41$, $p = .081$ (equivalence test), $\text{BF}_{01} = 5.52$. When we compared participants' ratings for the first and second context, we observed both the predicted increase in perceived pleasantness for CSs that were first paired with negative and then with positive USs, $\Delta M = 2.41$, 95% CI [1.64, ∞], $t(115.91) = 5.15$, $p < .001$, $\text{BF}_{10} = 3.50 \times 10^4$ (one-tailed), and the predicted decrease for CSs that were first paired with positive and then with negative USs, $\Delta M = 2.39$, 95% CI [1.61, ∞], $t(115.91) = 5.10$, $p < .001$, $\text{BF}_{10} = 5.90 \times 10^3$ (one-tailed). Moreover, we found an EC effect for the first context, $\Delta M = 2.19$, 95% CI [1.16, ∞], $t(103.46) = 3.53$, $p < .001$, $\text{BF}_{10} = 31.21$ (one-tailed), and a reversed EC effect for the second context, $\Delta M = 2.62$, 95% CI [1.59, ∞], $t(103.46) = 4.22$, $p < .001$, $\text{BF}_{10} = 153.33$ (one-tailed). Participants' prior knowledge about CSs did not affect these results, $F(1.17, 41.1) = 0.03$, $MSE = 10.61$, $p = .898$, $\hat{\eta}^2_G = .000$, $\text{BF}_{01} = 5.51$. Thus, although we observed no EC effect when we asked participants to report CS pleasantness in a new context at the end of the experiment, referring to and reinstating the learning contexts revealed changes in CS pleasantness throughout the learning procedure.

Discussion

The results of our counterconditioning experiment confirm the predictions derived from the temporal integration hypothesis and our simulation. First, we found the predicted expectancy-liking dissociation: Participants reported marked US expectancies throughout and, critically, at the end of the learning procedure—they expected CSs to appear with the most recently paired USs. In contrast, when participants provided CS pleasantness judgments immediately after the completion of the learning procedure and without reference to learning contexts, we found no EC effect. Second, participants made the predicted contextualized CS pleasantness judgments: We observed an EC effect for the initial acquisition context and a reversed EC effect for the counterconditioning context. These momentary CS pleasantness judgments reflected the changes in CS-US contingencies and corresponded to the intermittent US expectancy ratings. Hence, eliciting nondefault momentary evaluative judgments eliminated the expectancy-liking dissociation.

The temporal integration hypothesis posits that repeated judgments affect the adopted judgment strategy but do not affect evaluative learning. Research on contingency learning has shown that nondefault integrative contingency judgments can be elicited after completion of the learning procedure (Collins & Shanks, 2002; Matute et al., 2002). Our findings extend

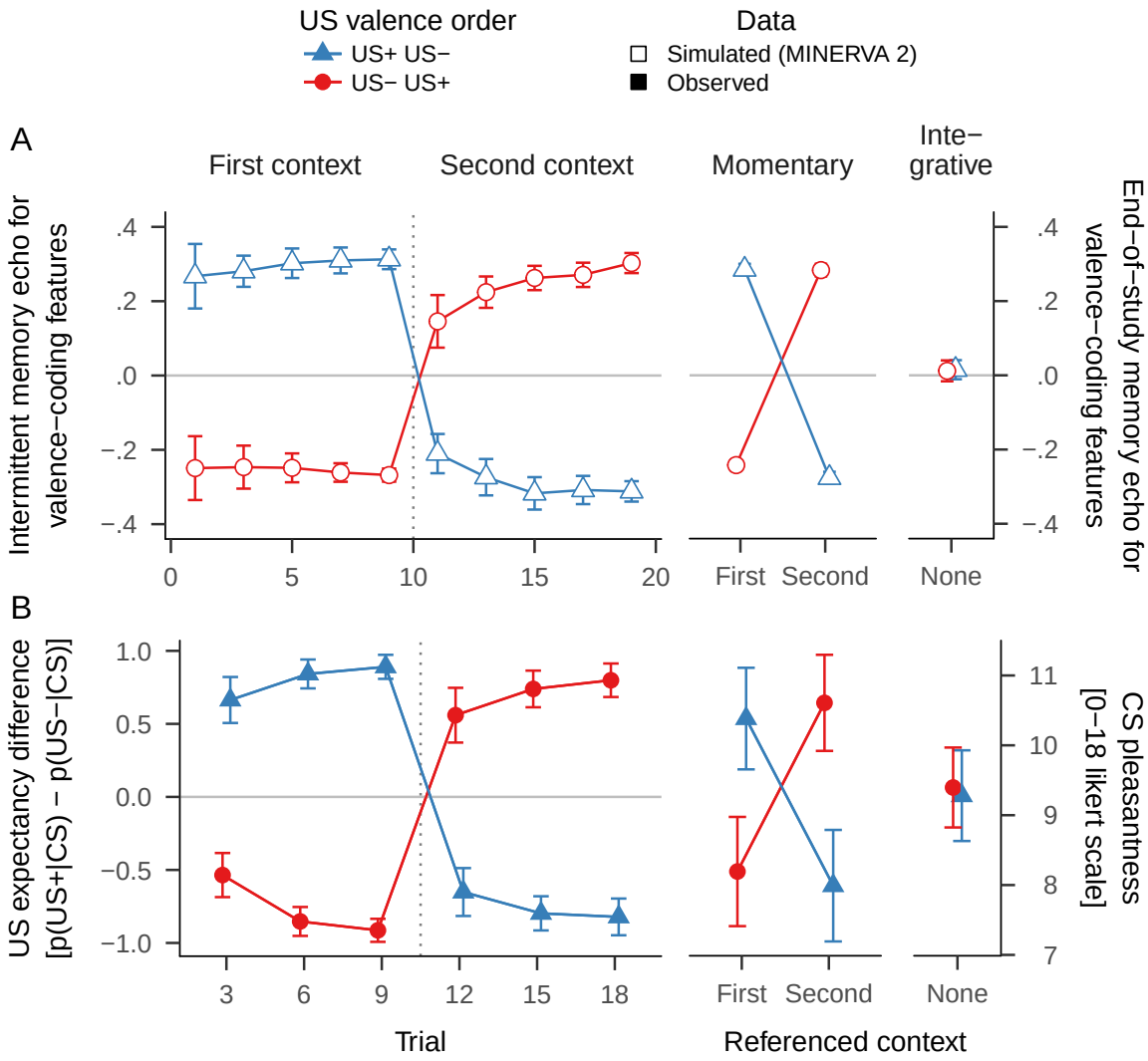

Figure 1. Simulated and observed US expectancy and CS pleasantness ratings for Experiment 1. Blue triangles indicate CSs paired with positive USs in the first and negative USs in the second context; red circles indicate CSs paired with negative USs in the first and with positive USs in the second context. **A** Mean normalized memory echo of valence-coding features predicted by MINERVA 2 indicative of the overall valence of the retrieved memory contents. The left plot shows valence retrieved during the learning procedure, the right plot shows the valence retrieved after completion of the learning procedure. Error bars represent 95% confidence intervals. **B** The left plot shows observed differences in mean US expectancy during the learning procedure. Positive values indicate expectancy for positive USs, negative values indicate expectancy for negative USs. The right plot shows observed mean CS pleasantness ratings after completion of the learning procedure. Error bars represent 95% within-subject confidence intervals; CS = Conditioned stimulus, US = Unconditioned stimulus.

these conclusions to CS pleasantness judgments. In this experiment, participants rated CS pleasantness only after completion of the learning procedure—they made no CS pleasantness judgments during the learning procedure. This approach is an improvement over previous studies in which CS pleasantness was assessed repeatedly during the learning procedure (e.g., Blechert, Michael, Williams, Purkis, & Wilhelm, 2008; Lipp et al., 2010; Lipp & Purkis, 2006) because it precludes that intermittent CS pleasantness judgments affected the evaluative learning process.

While our findings corroborate the dissociability of US expectancy and CS liking, they raise questions about the common dual-process interpretation of the expectancy-liking dissociation. The finding that US expectancy extinguishes while EC is resistant to extinction is commonly interpreted as evidence for a second associative learning process . In contrast, MINERVA 2 instantiates a candidate process-model of the single-process learning account (Mitchell et al., 2009). Drawing on the additional assumptions proposed by the temporal integration hypothesis (Lipp et al., 2010), the simulation illustrates that MINERVA 2 can predict the observed expectancy-liking dissociation in counterconditioning. Hence, absence of EC effects despite US expectancy can be explained by a single learning process.

Taken together, our findings support the assumptions of the temporal integration hypothesis that EC yields a single representation of CS-US pairings that informs both US expectancy and CS liking, and that their dissociation is caused by different default judgment strategies.

Experiment 2

The expectancy-liking dissociation reported in extinction procedures (e.g., Lipp & Purkis, 2006; Hermans et al., 2002) is the reverse of the dissociative pattern in the counterconditioning procedure: At the end of the learning procedure, participants no longer express US expectancies, but still exhibit an EC effect. As in Experiment 1, our reasoning was that inducing nondefault momentary judgments of CS pleasantness, by referring to and reinstating the learning contexts, would reveal extinction of EC effects.

We again began by simulating a simplified acquisition and an extinction procedure using MINERVA 2. The simulation method and assumptions were the same as for Experiment 1. In the extinction procedure, we paired CSs with USs in the first but presented them alone in the second context, Table 1. Conversely, in the acquisition procedure, we presented CSs alone in the first, and subsequently paired them with USs in the second context (see De Houwer, Baeyens, Vansteenwegen, & Eelen, 2000 for a similar approach). The CS-alone trials in the first context served to equate the number of CS-US pairings and valenced stimuli in the two contexts and, thus, to avoid attentional disengagement or mood effects.

The simulation predicted a pattern of results in line with the temporal integration hypothesis (Figure 2A): During the learning procedure, the valence of the retrieved memory contents closely followed the CS-US contingencies. In the acquisition procedure, the recalled memory contents remained neutral during the initial CS-alone trials but quickly acquired the USs' valence during the subsequent CS-US pairing trials. Conversely, in the extinction procedure, the recalled memory contents acquired the USs' valence during CS-US pairing

trials but quickly returned to a neutral baseline as a result of the combined context change and CS-alone trials. Thus, for the last trial of the acquisition procedure the simulation predicted expectations of the USs that had last been paired with CSs; in the last trial of the extinction procedure the simulation predicted the absence of a US expectancies.

The same pattern was predicted for end-of-study pleasantness judgments when the learning contexts were reinstated. For example, when a CS in the extinction procedure was presented in the first context—the context in which it was paired with a positive US—the valence of the retrieved memory contents was positive. However, when the same CS was presented in the second context—the context of CS-alone trials—the valence of the retrieved memory contents was neutral. In the new context—when no learning context was reinstated—the valence of the retrieved memory contents was comparable for CSs in the acquisition and extinction procedure. Furthermore, in both procedures the valence of the retrieved memory contents in the new context was comparable to that for the CS-US pairing context[4]. Hence, the simulation predicted comparable EC effects for default integrative end-of-study pleasantness ratings in the acquisition and extinction procedures. It also predicted comparable EC effects for momentary end-of-study pleasantness ratings in the CS-US pairing context and integrative ratings in the new context. Note that these predictions pertain to the current experimental design. As illustrated by the faint symbols in Figure 2A, the predictions differ for a design without a concurrent acquisition procedure or neutral CS-US pairs. We will return to this point in the General discussion.

In sum, the simulation predicted momentary judgments in intermittent US expectancy ratings; the momentary or integrative nature of judgments in end-of-study CS pleasantness ratings depended on the choice of context cues. Thus, our single-process memory model simulation predicted the well known expectancy-liking dissociation in extinction: No US expectancy in the last trial of the learning procedure but an EC effect in end-of-study CS pleasantness ratings without reference to the learning contexts. It furthermore predicted that extinction of EC could nonetheless be demonstrated in momentary end-of-study CS pleasantness ratings by referencing and reinstating the context of CS-alone trials. Hence, no expectancy-liking dissociation is expected when comparing momentary US expectancy to momentary CS pleasantness ratings. These predictions are in line with the explanation of the expectancy-liking dissociation proposed by the temporal integration hypothesis.

Based on the methodology of Experiment 1, we designed an experiment to test these predictions. We showed participants a stream of pictures in which CSs were either presented alone and subsequently paired with valent USs (acquisition procedure) or, conversely,

---

[4]Subsequent exploration identified two procedural factors that, in conjunction with the similarity-based retrieval mechanism of MINERVA 2, contribute to this prediction: (1) In the CS-US pairing context, the valence of the retrieved memory contents decreases as the number of neutral stimulus presentations in that context increases. This is because the retrieval cue for the CS-US pairing trials encompasses context features and, thus, to some degree activates all memory trace from that context (interference). (2) Conversely in the new context, the valence of the retrieved memory contents increases relative to the CS-US pairing context because the interference from neutral stimulus presentations is decreased. The attenuated interference is a consequence of the unique features of the new context, the nonlinear relationship between probe-trace similarity and trace activation, as well as the normalization of the echo contents.

paired with USs and subsequently presented alone (extinction procedure; see Table 1). We expected (1) to observe the predicted expectancy-liking dissociation between intermittent US expectancy ratings in the last trial of the learning procedure and end-of-study CS pleasantness ratings without reference to the learning contexts, but (2) to eliminate the expectancy-liking dissociation by demonstrating extinction of EC in momentary end-of-study CS pleasantness ratings when learning contexts are referenced and reinstated.

Methods

The experimental method, data analysis plan, and the following hypotheses were preregistered (https://osf.io/vnmby/registrations/): In the extinction procedure, we predicted (1) the expectancy for the US that had been paired with a given CS to extinguish towards the last trial, whereas we predicted (2) an EC effect in end-of-study CSs pleasantness ratings in the new context (i.e. resistance to extinction in integrative CS pleasantness judgments). Furthermore, we predicted a reduced EC effect for end-of-study CS pleasantness ratings in the context of CS-alone trials, compared to (3) the new context and (4) the context of CS-US pairing trials. We, additionally, wanted to test whether (5) EC effects in the new context were comparable to those for the CS-US pairing context. The last hypothesis was introduced to investigate the meta-analytical finding that the extinction of EC in default integrative end-of-study judgments exists but is small (Hofmann et al., 2010). Experimental design, materials, and procedure followed those of Experiment 1 except for the following changes.

Participants. To maximize the efficiency and informativeness of our study, we performed a sequential Bayesian analysis while the data were being collected (Rouder, 2014). Thus, the number of participants was not fixed a priori. We set a minimum sample size of $n = 20$ (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015) and planned to collect data until they provided strong evidence ($BF_{10} > 10$ or $BF_{01} > 10$) for or against our five hypotheses of primary interest or until we ran out of money (800€). We calculate Bayes factors at the end of each day of data collection.

We recruited 55 new participants. As preregistered we excluded one participant who performed the category recognition task at chance level, that is, they responded correctly to 25% or less of all category recognition questions. We assumed that at-chance category recognition is indicative of inattention because we instructed participants specifically to attend to pairings and detect regularities. We excluded one additional participant with a severe vision impairment, who was allowed to participate to obtain course credit. Thus, we stopped collecting data after 53 valid participants. At this point the data provided strong evidence for hypotheses 1-4. We deviated from our preregistered sampling plan and stopped data collection before the data provided an informative test of hypothesis 5 because it was not relevant to our theoretical predictions. The results of our Bayesian analysis are not affected by the premature termination of the data collection (Rouder, 2014). Participants' mean age was 22.24 years ($SD = 3.34$), 39 were female, 8 studied psychology or media psychology, all participants declared intact color vision, and 19 reported to have prior knowledge about the CS pictures.

Material. We adapted the 3-AFC US category recognition response format to the extinction procedure: Because the procedure included CS-alone trials, we added an additional "nothing" response option by which participants could indicate that a CS had not been paired with a US.

Procedure and design. For each participant we randomly assigned a positive, neutral, or negative US to each CS; the CS-US pair was then randomly assigned to the acquisition or extinction procedure. CSs in the acquisition procedure were presented alone in the first context and paired with USs in the second context. Conversely, in the extinction procedure, CSs were paired with USs in the first context and presented alone in the second context.

Instructions and assessment of our dependent measures were the same as in Experiment 1. However, we did not assess US identity recognition if CSs had been presented alone in a given context. Because each CS was paired with a US in only one of the two contexts, participants provided 18 US identity recognition responses, yielding 3 per experimental condition.

On average, participants took 51.15 minutes ($SD = 7.04$) to complete the study.

Results

Preregistered analyses (labeled confirmatory) will be followed by additional (exploratory) analyses. See Appendix B for analyses of participants' CS-US pairing memory.

US expectancy.

Confirmatory analyses.

We analyzed expectancies of the correct US using a 3 (*Valence*: Positive vs. Neutral vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) × 2 (*Context*: First vs. Second) × 3 (*Pairings*: 3 vs. 6. vs. 9) repeated-measures ANOVA. As in Experiment 1, participants quickly learned the CS-US contingencies, Figure 2B. As predicted, we found strong evidence that changes in expectancy of the correct US category across referenced contexts differed between acquisition and extinction procedures, $BF_{10} = 2.85 \times 10^{182}$. We observed this pattern irrespective of the valence of the US, $BF_{01} = 23.05$. Planned contrasts indicated that, averaged across US valences, expectancy for the correct US category after the ninth pairing increased from the first to the second context in the acquisition procedure ($M = 0.85$ 95% HDI [0.76, 0.94], $BF_{10} = 7.29 \times 10^{21}$, one-tailed) but decreased in the extinction procedure, $M = -0.67$ 95% HDI [−0.80, −0.55], $BF_{10} = 5.08 \times 10^{12}$, one-tailed. The data provided no noteworthy evidence as to whether participants had a residual expectancy for the correct US category at the end of the extinction procedure, $M = 0.06$ 95% HDI [0.00, 0.12], $BF_{01} = 1.14$, one-tailed. In sum, following CS-US pairings participants expected the correct US but US expectancy declined rapidly when CSs were subsequently presented alone.

Additionally, learning of CS-US contingencies proceeded faster in the second than the first context ($BF_{10} = 27.50$), regardless of US category valence ($BF_{01} = 50.74$) perhaps due

to familiarization with the learning procedure. We found no noteworthy evidence for any other effects of our experimental manipulations, all $BF_{01} \geq 1.87$.

Exploratory analyses.

Although participants may have retained some expectancy of the correct US at the end of the extinction procedure, this expectancy was markedly higher in the acquisition than in the extinction procedure, $M = 0.71$ 95% HDI [0.61, 0.82], $BF_{10} = 1.06 \times 10^{16}$ (one-tailed).

CS pleasantness.

Confirmatory analyses.

As a measure of the EC effect, we calculated difference scores between mean evaluative ratings of CSs that were paired with positive and negative USs ($\bar{x}_{\text{EC}} = \bar{x}_{\text{US+}} - \bar{x}_{\text{US−}}$) for every participant in every cell of the experimental design. We analyzed EC effects using a 2 (*Learning procedure*: Acquisition vs. Extinction) × 3 (*Referenced context*: First vs. Second vs. None) repeated-measures ANOVA. As predicted, referring to and reinstating learning contexts affected the EC effect differently depending on the learning procedure, $BF_{10} = 360.74$, Figure 2B. Planned contrasts indicated that when participants rated CS pleasantness in the new context, we found strong evidence for an EC effect in the extinction procedure, $M = 2.42$ 95% HDI [1.38, 3.50], $BF_{10} = 1.79 \times 10^3$ (one-tailed). Moreover, we found evidence that this EC effect was comparable to the EC effect in the acquisition procedure, $M = 0.32$ 95% HDI [0.00, 1.14], $BF_{01} = 12.30$ (one-tailed). When we compared EC effects for the first and second context, we observed both the predicted increase in the acquisition, $BF_{10} = 1.79 \times 10^5$ (one-tailed), as well as the predicted decrease in the extinction procedure, $BF_{10} = 6.77 \times 10^3$ (one-tailed). Critically, in the extinction procedure, the EC effect was reduced when participants rated CSs for the context of CS-alone trials compared to the new context, $BF_{10} = 74.23$ (one-tailed). We found only relatively weak evidence indicating that our learning procedure may not have extinguished the EC effect completely, $M = 1.17$ 95% HDI [0.19, 2.09], $BF_{10} = 4.60$. The comparison between the EC effects for the context of CS-US pairing trials and the new context was inconclusive, $BF_{01} = 2.03$. Similarly, in the acquisition procedure, the EC effect was reduced when participants rated CSs for the context of CS-alone trials compared to the new context, $BF_{10} = 576.00$ (one-tailed). The comparison between the EC effect for the context of CS-US pairing trials and the new context was again inconclusive, $BF_{01} = 1.87$. In sum, we found that the EC effect appeared to be resistant to extinction when participants rated CS pleasantness in a new context after completion of the learning procedure; but we found a reduced EC effect when we referenced and reinstated the context in which CS had been presented alone.

Exploratory analyses.

We additionally compared the EC effects between learning procedures for each of the contexts. In the first context, participants exhibited a larger EC effect in the extinction than in the acquisition procedure, $M = 2.77$ 95% HDI [1.26, 4.29], $BF_{10} = 122.86$ (one-sided). In the second context, the comparison was inconclusive, $M = 1.14$ 95% HDI [0.01, 2.37], $BF_{01} = 1.34$ (one-sided). We found evidence indicating that participants' prior knowledge about CSs did not affect our findings, $BF_{01} = 8.27$.

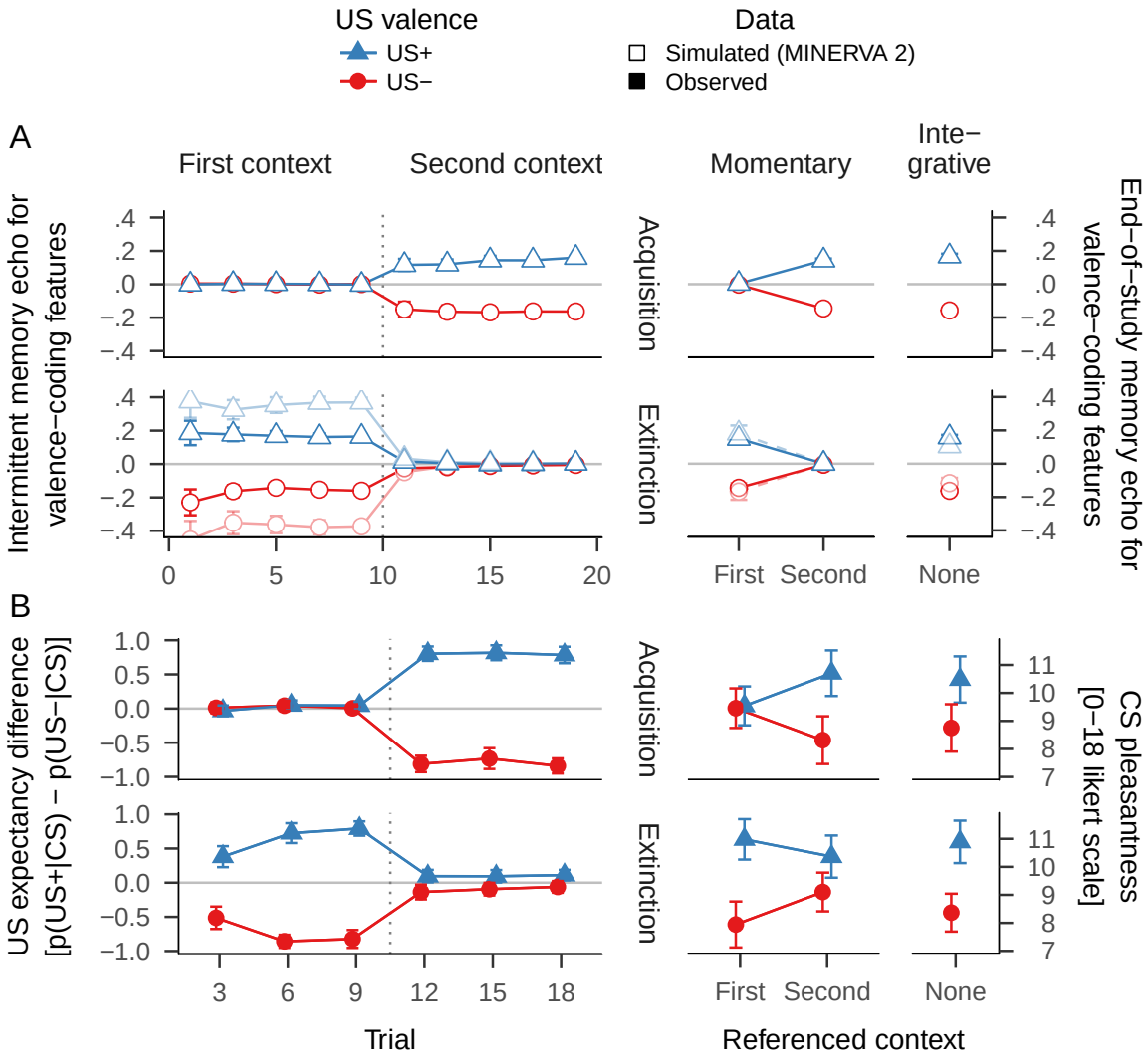

Figure 2. Simulated and observed US expectancy and CS pleasantness ratings for Experiment 2. Blue triangles indicate CSs paired with positive USs; red circles indicate CSs paired with negative USs. **A** Mean normalized memory echo of valence-coding features predicted by MINERVA 2 indicative of the overall valence of the retrieved memory contents. The left plot shows valence retrieved during the learning procedure, the right plot shows the valence retrieved after completion of the learning procedure. Faint symbols represent simulated ratings for a variant of our paradigm without the acquisition procedure or neutral CS-US trials. Error bars represent 95% confidence intervals. **B** The left plot shows observed differences in mean US expectancy during the learning procedure for acquisition (top) and extinction (bottom) procedures. Positive values indicate expectancy for positive USs, negative values indicate expectancy for negative USs. The right plot shows observed mean CS pleasantness ratings after completion of the learning procedure for acquisition (top) and extinction (bottom) procedures. Error bars represent 95% within-subject confidence intervals; CS = Conditioned stimulus, US = Unconditioned stimulus.

Discussion

Our results again confirm the predictions derived from the temporal integration hypothesis and our simulation. First, we replicated the expectancy-liking dissociation: In the last trial of the learning procedure, participants reported markedly higher US expectancies in the acquisition than in the extinction procedure. In contrast, when participants provided CS pleasantness judgments in a new context after the completion of the learning procedure (i.e., when learning contexts were not referenced), the EC effects did not differ between the acquisition and extinction procedure. Second, when we referenced and reinstated learning contexts participants again made contextualized CS pleasantness judgments: We observed extinction of the EC effect when participants evaluated CSs in the context of the CS-alone trials. These momentary CS pleasantness ratings, again, reflected the changes in CS-US contingencies and corresponded to the intermittent US expectancy ratings. Thus, as predicted, we elicited momentary CS pleasantness ratings and thereby eliminated the expectancy-liking dissociation. As in Experiment 1, this effect was obtained in the absence of potentially problematic intermittent CS pleasantness ratings. Jointly, our simulation and experimental findings provide further evidence that expectancy-liking dissociations can be explained as the result of different judgment strategies.

In Experiment 1 and 2, we assessed CS pleasantness repeatedly in different learning contexts. The repeated assessment may have introduced demand effects: Based on conversational norms, participants may have assumed that repeated ratings under varying conditions are expected to yield different responses. This also applies to intermittent US expectancy ratings. Moreover, these intermittent US expectancy ratings created a focus on CS-US pairings and US prediction. Although foci on pairings (e.g., Vansteenwegen et al., 2006; Fiedler & Unkelbach, 2011; Förderer & Unkelbach, 2012; Hu, Gawronski, & Balas, 2017; Vervliet, Vansteenwegen, Baeyens, Hermans, & Eelen, 2005) and on US prediction (e.g., Kattner & Green, 2015; Kattner, 2014; Zanon et al., 2012) are prevalent in EC research, it may limit the generalizability of our findings, and could be argued to impede automatic associative processes (Olson & Fazio, 2001). We addressed these limitations in Experiment 3.

Experiment 3

Our previous experiments indicate that the expectancy-liking dissociations in counterconditioning and extinction procedures are caused by different default judgment strategies and can be eliminated by inducing nondefault momentary CS pleasantness judgments. A comprehensive test of the temporal integration hypothesis, however, requires a concurrent manipulation of default judgment strategies for both US expectancy and CS pleasantness. The hypothesis predicts that, when judgment strategies are equated, US expectancy and CS pleasantness ratings should exhibit the same pattern of results. As a corollary, in an extinction procedure a comparison of nondefault integrative US expectancy and momentary CS pleasantness judgments should reveal a reversed expectancy-liking dissociation—extinction of EC effects despite continued US expectancy. Experiment 3 was a final test of the temporal integration hypothesis in which we concurrently manipulated the judgment strategies for US expectancy and CS pleasantness ratings.

As for CS pleasantness ratings, we assessed US expectancies only after completion of the learning procedure in a momentary fashion separately for each learning context, as well as in an integrative fashion in a new context. For this procedure, MINERVA 2's predictions match those of Experiment 2. Hence, we expected to observe (1) identical patterns for end-of-study US expectancy and CS pleasantness ratings and (2) a reversed expectancy-liking dissociation (Figure 2A).

Methods

The experimental method, data analysis plan, and the following hypotheses were preregistered (https://osf.io/vnmby/registrations/): In the extinction procedure, we predicted (1) persistent US expectancy in end-of-study judgments that referred to both learning contexts (i.e. resistance to extinction in integrative US expectancy judgments). Moreover, we predicted lower US expectancy ratings for the context of CS-alone trials when compared to (2) the context of CS-US pairing trials as well as (3) to both learning contexts. For CS pleasantness ratings, we predicted the same pattern: Despite the extinction procedure, we expected to observe (4) an EC effect in end-of-study judgments in the new context. Moreover, we expected (5) a reduced EC effect for end-of-study CS pleasantness ratings for the context of CS-alone trials compared to the new context and (6) the context of CS-US pairing trials.

Experimental design, materials, and procedure followed those of Experiment 2 except for the following changes.

Participants. As in Experiment 2, we performed a sequential Bayesian analysis with minimum sample size of $n = 20$ per between subject condition ($N = 120$). We set out to collect data until they provided strong evidence for or against our six hypotheses of primary interest or until we ran out of money (1920€).

We recruited 273 new participants. As preregistered, we excluded 17 participants who performed the category recognition task at chance level, 57 participants who performed poorly at the identification task during the learning procedure (below $Q_1 - 1.5 \times IQR$, i.e. at least one incorrect response; see Procedure), and one participant who aborted the experiment. Thus, we stopped collecting data after 202 valid participants. At this point the data provided strong evidence for hypotheses 1, 2, and 5. Participants' mean age was 23.61 years ($SD = 6.41$), 146 were female, and 32 studied psychology or media psychology. 7 participants reported vision impairments: five were red-green color blind, one had astigmatism and another had a blind eye. 74 participants reported to have prior knowledge about the CS pictures.

Material. In contrast to Experiment 2, we did not collect intermittent US expectancy ratings. Instead, we asked participants to categorize USs ("What do you see right now?") in a 4-AFC task as photographs of either humans, animals, or objects or to indicate that no US was presented. The categorization task served to engage participants during the learning procedure in a manner comparable with our previous experiments. Analogous to CS pleasantness ratings, participants judged US expectancy for each CS after completion of the learning procedure for different contexts. We instructed participants that they would repeat a few trials from the learning procedure. We presented only the CSs and asked "With what

probability would you expect a photograph of a human [animal/object] with this creature?" Previous studies have similarly assessed expectancy retrospectively by asking participants to graph the evolution of their US expectancy during the learning procedure (e.g., Raes, De Houwer, Verschuere, & De Raedt, 2011; Vansteenwegen et al., 2005; Vervliet et al., 2005). To elicit momentary judgments, we noted that these trials were drawn from the first (second) half of the experiment. To elicit integrative judgments, we instructed participants that the trials were "selected randomly from the first and second half of the experiment" with equal probability. We noted that CSs would be shown in the center on neutral background to obscure which half of the experiment the trials were drawn from.

Procedure and design. In each of the six subblocks of the learning procedure, we collected US categorization responses for one CS from every US valence (including neutral CS-US pairs) following the third presentation of the CS-US pair. We removed the CS, but the US—if a US had been presented—remained on screen until participants responded. We made this task deliberately easy to avoid drawing too much attention to USs and away from CSs during the learning procedure.

After completion of the learning procedure, participants rated US expectancy and CS pleasantness for each CS. In contrast to Experiment 2, the context for CS pleasantness and US expectancy ratings was manipulated between participants, i.e., participants rated each CSs only for one context. We, thus, collected 3 US expectancy and CS pleasantness ratings per experimental condition and 18 per participant. Additionally, we manipulated the order of US expectancy or CS pleasantness ratings to control for possible order effects (Heycke et al., 2017).

On average, participants took 49.62 minutes ($SD = 12.71$) to complete the study.

Results

Preregistered analyses (labeled confirmatory) will be presented first, followed by additional (exploratory) analyses. In addition to the analyses presented here, we repeated all analyses with a modified set of exclusion criteria: We included participants who made no more than one incorrect response in the intermittent US identification task—some participants reported accidentally clicking the wrong button—but excluded three participants who invariably used the scale mid-point in CS pleasantness ratings (see the online supplemental material). By and large, we found the same results; we indicate noteworthy changes in the exploratory analyses sections. See Appendix B for analyses of participants' CS-US pairing memory.

US expectancy.

Confirmatory analyses.

We analyzed expectancies of the correct US using a 3 (*Valence*: Positive vs. Neutral vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) × 3 (*Referenced context*: First vs. Second vs. Both) × 2 (*DV order*: CS pleasantness first vs. US expectancy first) ANOVA with repeated measurements on the first two factors. As predicted, we found strong evidence that the changes in US expectancy across contexts differed between acquisition and

extinction procedures, $BF_{10} = 1.03 \times 10^{33}$, Figure 3. We observed this pattern irrespective of US valence ($BF_{01} = 66.72$) and of whether US expectancy was assessed before or after CS pleasantness, $BF_{01} = 8.50$. We therefore analyzed all data and averaged across US valences.

As predicted, planned contrasts indicated that expectancy for the correct US category increased from the first to the second learning context in the acquisition procedure ($BF_{10} = 11.19$, one-tailed) but decreased in the extinction procedure, $BF_{10} = 6.12 \times 10^4$ (one-tailed). When we referenced both learning contexts, we found strong evidence that participants expected the correct USs despite the previous extinction procedure, $M = 0.37$ 95% HDI [0.26, 0.48], $BF_{10} = 1.18 \times 10^7$, one-tailed. The comparisons of US expectancy for both contexts versus the second context was inconclusive in both acquisition ($BF_{10} = 1.38$, one-tailed) and extinction procedures, $BF_{10} = 1.92$ (one-tailed). There was no noteworthy evidence to suggest that there was any other effect of our manipulations, $BF_{01} \geq 1.54$. In sum, participants' end-of-study US expectancies corresponded to CS-US contingencies when we referenced and reinstated the learning contexts.

Exploratory analyses.

Because we found no conclusive evidence for or against integrative judgments in the preregistered between-participant comparisons of ratings for the second and the new context, we additionally compared the differences between the acquisition and extinction procedures for all referenced contexts. For the first learning context, participants expressed higher expectancy for the correct US in the extinction than in the acquisition procedure, $M = 0.37$ 95% HDI [0.25, 0.48], $BF_{10} = 2.97 \times 10^6$ (one-sided). This pattern was reversed in the second context: Participants expressed higher expectancy for the correct US in the acquisition than in the extinction procedure, $M = 0.24$ 95% HDI [0.14, 0.33], $BF_{10} = 9.85 \times 10^3$ (one-sided). Critically, when we referenced both learning contexts we found some evidence that expectancy for the correct US did not differ between acquisition and extinction procedures, $M = 0.03$ 95% HDI [0.00, 0.07], $BF_{01} = 5.55$ (one-sided). These additional analyses indicate that, like the EC effect, US expectancy appeared to be resistant to extinction when we referenced both learning contexts. Hence, we conclude that we successfully elicited integrative US expectancy judgments.

Compared to the intermittent ratings in Experiments 1 and 2, participants reported expecting USs in the context of CS-alone trials and overall their expectancies were less pronounced. Further analyses suggested that this reflects memory confusions of the learning contexts.

CS pleasantness.

Confirmatory analyses.

We analyzed EC effects using a 2 (*Learning procedure*: Acquisition vs. Extinction) × 3 (*Referenced context*: First vs. Second vs. None) × 2 (*DV order*: CS pleasantness first vs. US expectancy first) ANOVA with repeated measurements on the first factor. As predicted, referring to and reinstating learning contexts affected the EC effect differently depending on the learning procedure, $BF_{10} = 1.45 \times 10^3$, Figure 3. This finding was not affected by the order of DVs ($BF_{01} = 5.30$) and, thus, we analyzed all data. End-of-study CS pleasantness ratings in the new context provided some evidence for an EC effect in the extinction

conditions, $M = 0.99$ 95% HDI [0.20, 1.80], $BF_{10} = 4.35$ (one-tailed). Moreover, we found evidence, albeit weak, that this EC effect was of comparable magnitude in the extinction and acquisition procedure, $M = 0.60$ 95% HDI [0.00, 1.52], $BF_{01} = 4.64$ (one-tailed). When we compared participants' CS pleasantness ratings for the first and second context, we observed both the predicted increase in the EC effect in the acquisition procedure, $BF_{10} = 19.44$ (one-tailed), as well as the predicted decrease in the extinction procedure, $BF_{10} = 38.71$ (one-tailed). In the extinction procedure, the EC effects for the context of CS-alone trials and the new context were of comparable magnitude , $BF_{01} = 5.10$ (one-tailed). EC in the context of CS-alone trials was not extinguished completely, $M = 1.31$ 95% HDI [0.34, 2.29], $BF_{10} = 7.19$; but we found evidence for partial extinction. The EC effect was clearly larger in the context of CS-US pairing trials than in the new context, $BF_{10} = 88.80$, in line with the meta-analytic finding. Similarly, in the acquisition procedure, the EC effect for the context of CS-US pairing trials was larger than for the new context, $BF_{10} = 10.95$. The comparison between the EC effect for the context of CS-alone trials and the new context was, however, inconclusive, $BF_{01} = 1.70$ (one-tailed). We found no noteworthy evidence for any other effects of our manipulations, $BF_{10} \leq 2.82$. In sum, we found some indication that EC effects were comparable in the acquisition and extinction procedures when participants rated CS pleasantness in the new context after completion of the learning procedure. We also observed the predicted extinction of EC in nondefault momentary CS pleasantness judgments: The EC effect was larger for the context of CS-US pairing trials than for the context of CS-alone trials.

Exploratory analyses.

Additionally, in the first learning context participants exhibited a larger EC effect in the extinction than in the acquisition procedure, $M = 2.24$ 95% HDI [0.80, 3.70], $BF_{10} = 22.69$ (one-sided). This pattern reversed in the second context: Participants exhibited a larger EC effect in the acquisition than in the extinction procedure, $M = 2.30$ 95% HDI [0.82, 3.68], $BF_{10} = 28.47$ (one-sided). The data were uninformative as to whether participants' prior knowledge about CSs affected these findings, $BF_{01} = 1.85$.

In the exploratory analysis using the modified set of exclusion criteria ($n = 229$), we found stronger evidence in support of an EC effect in the new context in the extinction procedure, $BF_{10} = 11.60$ (one-tailed). In this larger sample we also found stronger evidence indicating that the magnitude of this EC effect was comparable in the extinction and the acquisition procedure, $BF_{01} = 7.03$ (one-tailed).

Discussion

Experiment 3 replicated and extended our previous findings in the absence of both intermittent US expectancy ratings and repeated end-of-study assessment of US expectancy or CS pleasantness across different contexts. We successfully elicited momentary US expectancy judgments by referring to and reinstating the learning contexts that adequately reflected the CS-US contingency changes: In both learning procedures, participants' US expectancy was larger in the context of CS-US pairing trials than in the context of CS-alone trials. Following extinction learning, participants reported residual US expectancies in the context

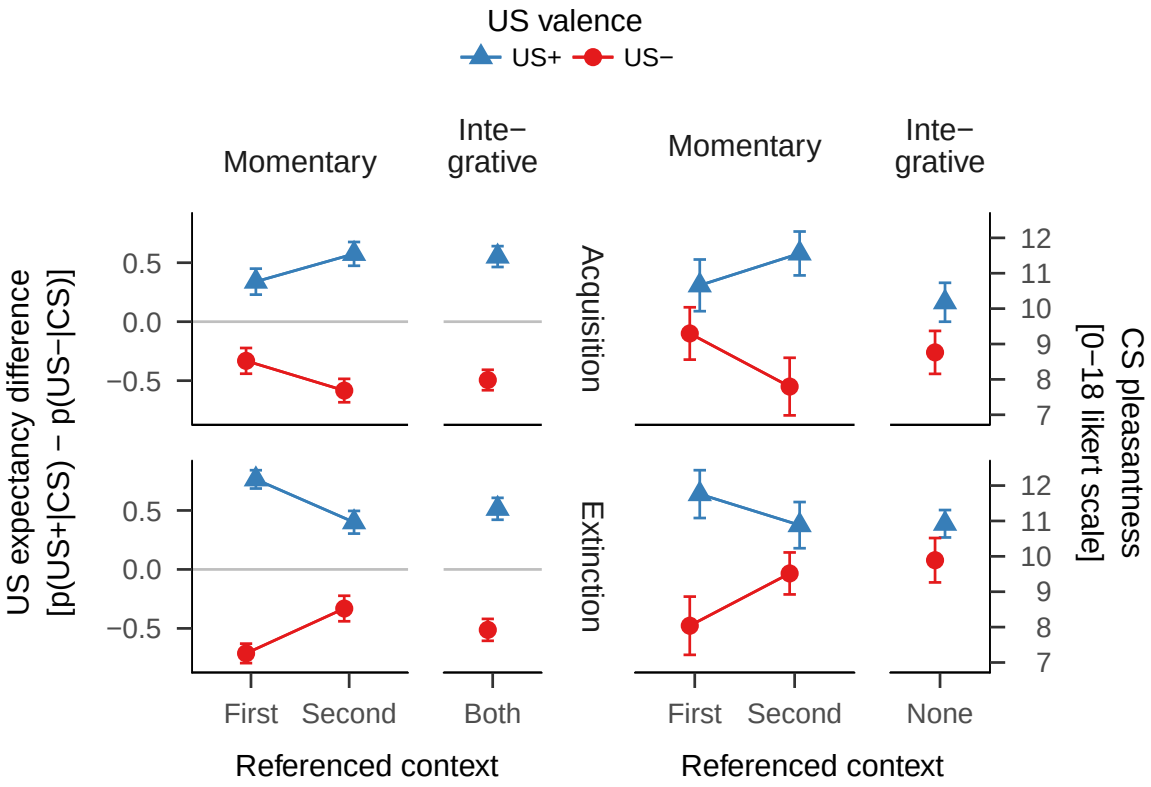

Figure 3. US expectancy and CS pleasantness ratings at the end of Experiment 3. The left plot shows observed differences in mean US expectancy for acquisition (top) and extinction (bottom) procedures. Positive values indicate expectancy for positive USs, negative values indicate expectancy for negative USs. The right plot shows observed mean CS pleasantness ratings after completion of the learning procedure for acquisition (top) and extinction (bottom) procedures. Error bars represent 95% within-subject confidence intervals; CS = Conditioned stimulus, US = Unconditioned stimulus.

of CS-alone trials, but US expectancy was higher in the acquisition procedure. Critically, US expectancy in the acquisition and extinction procedures was comparable when we referenced both learning contexts, reflecting a nondefault integrative judgment strategy.

As predicted by the temporal integration hypothesis and our simulation, CS pleasantness ratings exhibited the same pattern of results: In both learning procedures, EC effects were larger in the context of CS-US pairing trials than in the context of CS-alone trials. Following extinction learning, participants' ratings exhibited a residual EC effect in the context of CS-alone trials, but the effect was markedly higher in the acquisition procedure. Again, when no learning context was referenced the EC effect was comparable between acquisition and extinction procedures, reflecting the default integrative judgment strategy.

In summary, the extinction procedure showed the well-known expectancy-liking dissociation between default momentary US expectancy ratings and integrative CS pleasantness ratings. Conversely, we found the reversed dissociation between nondefault integrative US expectancy and momentary CS pleasantness ratings. Thus, after equating judgment

strategies, US expectancy and CS pleasantness exhibited the same pattern of results. We conclude that expectancy-liking dissociations can be accounted for by differences in default judgment strategies and do not necessitate two distinct learning systems.

General Discussion

To explain expectancy-liking dissociations in EC (e.g. Baeyens et al., 1988, 2005; Hermans et al., 2002) with a single learning process, Lipp et al. (2010) proposed that US expectancy and CS pleasantness ratings afford different default judgment strategies: US expectancy ratings reflect momentary whereas CS pleasantness ratings reflect integrative summaries of the learning history. We tested this temporal integration hypothesis by manipulating participants' judgment strategies after completion of the learning procedure in a counterconditioning and two extinction experiments. Under default conditions (i.e., momentary US expectancy and integrative CS pleasantness judgments), we replicated two expectancy-liking dissociations: Counterconditioning produced no EC effects although participants US expectancies reflected the contingencies in the second part of the learning phase; conversely, extinction produced EC effects in the absence of US expectancies. Our findings corroborate that these dissociations were caused by the difference in strategies. First, we eliminated these dissociations by equating the judgment strategies across measures: CS pleasantness ratings corresponded to the respective US expectancy ratings after we elicited (nondefault) momentary CS pleasantness judgments (i.e., by referring to and reinstating the context of the initial or opposed CS-US pairings in the counterconditioning procedure in Experiment 1, or the context in which CSs had been presented alone as in the extinction procedures of Experiments 2 and 3). Furthermore, we reversed the expectancy-liking dissociation in the extinction paradigm by contrasting (nondefault) integrative US expectancy with (nondefault) momentary CS pleasantness judgments. Results showed extinction of EC but resistance to extinction of US expectancy. Our findings demonstrate that the expectancy-liking dissociations reported in the literature can be produced as the result of different judgment strategies afforded by the dependent measures. Hence, contrary to previous interpretations, expectancy-liking dissociations do not necessitate a second learning process; they can be parsimoniously explained by a single learning process.

Amending and extending Lipp et al. (2010)'s temporal-integration hypothesis, we illustrate how the learning history can be conserved and utilized to perform judgment tasks: Based on previous theorizing (Mitchell et al., 2009), we instantiated learning and retrieval processes by the unitary episodic memory model MINERVA 2 (Hintzman, 1984, 1986, 1988), which enabled us to make specific predictions for both US expectancy and CS evaluations that were subsequently corroborated by our experiments.

Additional findings and limitations

Hofmann et al. (2010) found partial extinction of EC in studies that assessed the EC effect both after the acquisition and again after the extinction phase. At first glance, this finding may seem to contradict the results of our simulation (Figure 2A), which predicted comparable CS pleasantness ratings for intermittent ratings at the end of the acquisition

procedure, integrative end-of-study ratings, as well as momentary ratings in the CS-US pairing context, However, this discrepancy is due to procedural factors. For learning procedures that elicit momentary CS evaluations via repeated ratings, MINERVA 2 predicts partial extinction. In such cases, postextinction ratings should more strongly reflect recent CS-alone trials. Moreover, the model predictions also depend on other aspects of the experimental designs, such as the presentation of neutral stimuli (e.g., CS-alone trials or neutral CS-US pairs) during the initial acquisition phase in the present studies. This is because the common context causes activation of neutral stimuli, which in turn attenuate CS pleasantness ratings. As illustrated by the faint symbols in Figure 2A, MINERVA 2 predicts the partial extinction effect for designs without neutral stimuli during the initial acquisition phase.

In the present studies, we manipulated learning contexts overtly. However, we believe that our findings also pertain to evaluative learning without context manipulations. Lipp and Purkis (2006) found that using paper and pencil rather than a computer for end-of-study valence assessment is sufficient to elicit integrative CS pleasantness judgments without any explicit context manipulation. Similar renewal effects have been found in social impression formation (AAB renewal; Gawronski, Rydell, Vervliet, & De Houwer, 2010, Experiment 4). Hence, the standard end-of-study evaluation assessment may act as context change, affect judgment strategies, and produce renewal effects. Moreover, we assume that in the absence of explicitly induced context changes, participants spontaneously generate and use temporal contexts to structure the incoming information and that these contexts can affect behavior (Matute et al., 2011; Zacks et al., 2007). Due to the use of external contexts, contextualization in our study may have been more pronounced but—we assume—not qualitatively different from previous studies. We plan to test these assumptions in future research.

Yet another type of context has been employed in studies on feature-positive learning in EC—another procedure that has produced an expectancy-liking dissociation (Baeyens, Crombez, De Houwer, & Eelen, 1996; Baeyens, Hendrickx, Crombez, & Hermans, 1998). In this paradigm, a CS was paired with a negative US only in the presence of (or subsequent to) a feature stimulus; in its absence, the CS was presented alone. EC effects were obtained both when the CS was rated in the presence and absence of the feature stimulus—even when participants correctly reported the stimulus contingencies. For such a procedure, MINERVA 2 predicts a reduced, albeit nonzero, EC effect when the CS is rated in the absence compared to the presence of the feature stimulus. If the nonsignificant reduction, which has so far been obtained only with relatively small samples (Baeyens et al., 1996, 1998), proves robust in high-powered studies, it is a finding that challenges our model and may necessitate further refinement.

Comparing our findings to those from studies on social impression formation also reveals a potentially interesting inconsistency (for a review see Gawronski et al., 2018). In their counterconditioning-like paradigm, Gawronski et al. (2010) found that participants' evaluations in a new context reflected the valence of the initially presented information (ABC renewal), whereas the present studies instead found neutral evaluations. Interestingly, Gawronski et al. (2018) also failed to observe those renewal effects in preliminary EC

studies (p. 43). Although social impression formation and EC procedures differ in many aspects, we speculate that attentional processes may explain the contradictory results. Our model assumed constant attention to context—we informed participants that the learning procedure consisted of two phases. If, instead, we assume that attention to context increases in the second context—as do Gawronski et al. (2018)—our model predicts ABC renewal. Conversely, when Gawronski et al. (2010) enhanced attention to the first context, ABC renewal was eliminated (Experiment 4). The effects of attention to, and encoding of, context features is closely linked to the above considerations about the different types of context manipulations and deserves further study.

Implications

The role of dependent measures. We demonstrate the expectancy-liking dissociation in extinction learning—as well as its reversal—while holding encoding constant and manipulating only the retrieval or judgment process. Somewhat relatedly, Gawronski et al. (2014) interpreted the meta-analytical finding of a small reduction of EC due to extinction (Hofmann et al., 2010) as an artifact of judgment-related nuisance processes. They argued that extinction procedures do not affect the underlying evaluative representations. Our results corroborate the importance of judgment processes in evaluative responses, but they also highlight that similar judgment processes affect expectancy judgments. With this in mind, the expectancy-liking dissociations and the resistance to extinction of EC can similarly be construed as an artifact of judgment-related processes. Resistance to extinction appears to reflect different judgment strategies rather than characteristics of separable learning systems.

More generally, our findings illustrate that conclusions about latent processes require a good understanding, and careful experimental treatment, of the dependent measures. Dissociations taken to imply the operation of different learning processes may alternatively be explained by differences in retrieval or performance processes that bear on the assessed variable. Without a good understanding of the dependent measures (i.e., without an established measurement theory), contrasting these measures runs the risk of comparing apples and oranges. Instead, stronger and more direct tests of dual-process claims can be achieved if the outcomes of experimental manipulations that selectively target the two postulated processes are assessed and compared on a single dependent variable.

Explicit versus implicit measures. Another dissociation often discussed in the EC literature is the one between direct measures, such as ratings, and indirect measures, such as evaluative priming. In research on extinction, these dissociations have often been interpreted as evidence for dual-processes theories (e.g., Gawronski et al., 2014; Kattner & Green, 2015). This interpretation rests on the assumption that ratings primarily reflect explicit learning and priming measures reflect implicit learning. Challenging this assumption, EC studies have routinely used direct measures to assess implicit learning (e.g., Olson & Fazio, 2001; Hütter, Sweldens, Stahl, Unkelbach, & Klauer, 2012); and effects supporting explicit learning have been obtained on indirect measures (e.g., EC requires awareness; Pleyers, Corneille, Luminet, & Yzerbyt, 2007; Stahl, Unkelbach, & Corneille, 2009). A recent review of implicit-explicit dissociations in attitude learning concludes that there is

little evidence for the assumption of a link between learning mode and expression mode (Corneille & Stahl, n.d.). Taken together, these findings illustrate that dissociations between direct and indirect measures are often absent; and where obtained, they would merely be consistent with dual-process assumptions but fail to corroborate them.

Gawronski et al. (2014) reported that extinction reduced EC effects on the direct evaluative ratings but did not affect the indirect evaluative priming measure. According to the temporal integration hypothesis, the extinction of EC on the direct measure reflects its context-sensitivity: When repeatedly asked to evaluate the CSs following acquisition and again after extinction, this repetition induces a momentary judgment strategy—participants' latter ratings reflected primarily the information encoded during the extinction phase. Assuming that the resistance to extinction inferred from evaluative priming is not merely an artifact of the measure's inferior reliability and sensitivity, a possible explanation is that judgment strategies can be adapted more readily for explicit ratings than in evaluative priming. In contrast to direct expressions of attitudes, context-sensitive responses require considerable effort in evaluative priming measures (on top of task-specific knowledge and strategies; Klauer & Teige-Mocigemba, 2007; Teige-Mocigemba & Klauer, 2008). Thus, a more targeted manipulation may be necessary to modify the default integrative judgment strategy in evaluative priming. The research reviewed by Gawronski et al. (2018) suggests momentary judgments can be elicited in indirect measures by introducing context manipulations similar to ours. To the degree that such context effects indeed affect indirect measures, our findings should generalize to these measures.

Our aim was not to conclusively rule out the existence of a second learning process. Specific conditions may, for example, allow for an additional implicit misattribution (IM) of unconditioned evaluative responses to the paired CSs (Olson & Fazio, 2001). In contrast to our stimulus-stimulus (S-S) learning account, IM postulates stimulus-response (S-R) learning—links between CSs and evaluative responses. Other properties of CS-US pairings, such as context, should be inconsequential for IM; it therefore cannot readily explain the contextualized EC effects we observed. Our procedure realized some conditions that are taken to promote IM (a seemingly random stream of stimuli and incidental learning of US valence; Hütter & Sweldens, 2013; Jones et al., 2009) but it could be adapted to further bolster misattribution of evaluative responses. Specifically, incidental instructions, together with simultaneous onset of CS and US (Hütter & Sweldens, 2013) may reveal context-insensitive IM.

Memory models that account for learning phenomena. Traditionally, learning and memory have often been studied by separate research communities. However, as illustrated by Tolman's notion of memory as latent learning, it has been clear that these phenomena are overlapping as the effects studied in learning research must be mediated by (some form of) memory. The present study is one of several that sketch how to integrate learning phenomena into established memory theorizing. Here we recast EC—traditionally construed as a learning phenomenon—in terms of episodic memory theory. We view evaluative learning as encoding and retrieval of episodic knowledge that may later be used to construct adaptive judgments.

Jamieson, Crump, and Hannah (2012) have proposed a related account of associative

learning based on an adapted memory model called Minerva-AL. The crucial difference between MINERVA 2 and Minerva-AL resides in the encoding mechanism: Instead of passively encoding episodes, Minerva-AL assumes that CSs evoke predictions about USs and that only discrepancies between predictions and observed events—the prediction error—is stored in memory. Although Minerva-AL makes the same representational assumptions and posits the same retrieval mechanisms as MINERVA 2, the discrepancy-encoding mechanism results in a model that is closely related to classical learning models such as the Rescorla-Wagner model (Miller, Barnet, & Grahame, 1995; Rescorla & Wagner, 1972; Siegel & Allan, 1996).

Minerva-AL was developed to account for phenomena in classical conditioning, which is believed to be a highly intentional learning procedure in which outcome expectations drive responses (Mitchell et al., 2009). Such conditions may encourage and even require continuous predictions and error monitoring. In EC, however, incidental paradigms, which obfuscate CS-US contingencies, are of particular relevance to the single- vs. dual-process debate (Corneille & Stahl, n.d.; Olson & Fazio, 2001; Stahl & Heycke, 2016). It is unclear whether and how participants generate and test predictions about CS-US associations in incidental paradigms; passive encoding of CS-US pairings, as assumed by MINERVA 2, may be a more appropriate assumption here. While MINERVA 2 predicted our findings despite the intentional learning instructions and intermittent US expectancy ratings, extending our approach to other paradigms and effects may necessitate modifications or additional assumptions. Comparing MINERVA 2 and Minerva-AL and exploring their limitations with respect to EC and classical conditioning is an interesting direction for future research.

It has long been known that MINERVA 2 requires additional assumptions to account for some of the critical empirical findings in recognition memory. For example, recall strategies have to be assumed to account for some findings in associative recognition (Clark & Gronlund, 1996). Yet, MINERVA 2, and the class of global-matching models to which it belongs, have been influential. The fact that the model predicted the outcomes of our experiments is an encouraging first indication that MINERVA 2 and related memory models (Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989; Kelly, Mewhort, & West, 2017; Shiffrin & Steyvers, 1997) are viable candidates for process models of EC that merit further exploration.

References

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Baeyens, F., Crombez, G., De Houwer, J., & Eelen, P. (1996). No Evidence for Modulation of Evaluative Flavor in Humans. *Learning and Motivation*, *27*(2), 200–241. https://doi.org/10.1006/lmot.1996.0012

Baeyens, F., Crombez, G., Van den Bergh, O., & Eelen, P. (1988). Once in contact always in contact: Evaluative conditioning is resistant to extinction. *Advances in Behaviour Research and Therapy*, *10*(4), 179–199. https://doi.org/10.1016/0146-6402(88)90014-8

Baeyens, F., & De Houwer, J. (1995). Evaluative conditioning is a qualitatively distinct form of classical conditioning: A reply to Davey (1994). *Behaviour Research and Therapy*, *33*(7), 825–831. https://doi.org/10.1016/0005-7967(95)00021-O

Baeyens, F., Dı'az, E., & Ruiz, G. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cognition and Emotion*, *19*(2), 245–268. https://doi.org/10.1080/02699930441000300

Baeyens, F., Hendrickx, H., Crombez, G., & Hermans, D. (1998). Neither Extended Sequential nor Simultaneous Feature Positive Training Result in Modulation of Evaluative Flavor in Humans. *Appetite*, *31*(2), 185–204. https://doi.org/10.1006/appe.1998.0167

Baeyens, F., Vansteenwegen, D., & Hermans, D. (2009). Associative learning requires associations, not propositions. *Behavioral and Brain Sciences*, *32*(02), 198. https://doi.org/10.1017/S0140525X09000867

Biegler, P., & Vargas, P. (2013). Ban the Sunset? Nonpropositional Content and Regulation of Pharmaceutical Advertising. *The American Journal of Bioethics*, *13*(5), 3–13. https://doi.org/10.1080/15265161.2013.776127

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*(1), 58–70. https://doi.org/10.1016/j.lmot.2007.03.003

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*(1), 37–60. https://doi.org/10.3758/BF03210740

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition*, *30*(7), 1138–1147. https://doi.org/10.3758/BF03194331

Corneille, O., & Stahl, C. (n.d.). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*.

De Houwer, J. (1998). Leren is eenvoudig, doen is moeilijk: Een nieuwe visie op het onderscheid tussen evaluatieve en Pavloviaanse conditionering [Learning is simple, performance is difficult: A new view on the relation between evaluative and Pavlovian conditioning]. *Gedragstherapie*, *31*, 49–66.

De Houwer, J. (2011). Evaluative conditioning: A review of procedure knowledge and mental process theories. In T. R. Schachtman & S. Reilly (Eds.), *Applications of learning and conditioning*. Oxford, UK: Oxford University Press.

De Houwer, J., Baeyens, F., Vansteenwegen, D., & Eelen, P. (2000). Evaluative conditioning in the picturepicture paradigm with random assignment of conditioned stimuli to unconditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *26*(2), 237–242. https://doi.org/10.1037/0097-7403.26.2.237

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological*

Bulletin, *127*(6), 853–869. https://doi.org/10.1037/0033-2909.127.6.853

Dwyer, D. M., Jarratt, F., & Dick, K. (2007). Evaluative conditioning with foods as CSs and body shapes as USs: No evidence for sex differences, extinction, or overshadowing. *Cognition and Emotion*, *21*(2), 281–299. https://doi.org/10.1080/02699930600551592

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion*, *25*(4), 639–656. https://doi.org/10.1080/02699931.2010.513497

Förderer, S., & Unkelbach, C. (2012). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS relations. *Cognition and Emotion*, *26*(3), 534–540. https://doi.org/10.1080/02699931.2011.588687

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Gast, A., & De Houwer, J. (2014). Is evaluative conditioning really resistant to extinction? Evidence for changes in evaluative judgements without changes in evaluative representations. *Cognition and Emotion*, 1–15. https://doi.org/10.1080/02699931.2014.947919

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized Attitude Change. In J. M. Olson (Ed.), *Advances in Experimental Social Psychology* (Vol. 57). Academic Press. https://doi.org/10.1016/bs.aesp.2017.06.001

Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General*, *139*(4), 683–701. https://doi.org/10.1037/a0020315

Hermans, D., Crombez, G., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2002). Expectancy-learning and evaluative learning in human classical conditioning: Differential effects of extinction. In S. P. Shohov (Ed.), *Advances in psychology research* (Vol. 12, pp. 17–40). Hauppauge, NY: Nova Science Publishers.

Heycke, T., Aust, F., & Stahl, C. (2017). Subliminal influence on preferences? A test of evaluative conditioning for brief visual conditioned stimuli using auditory unconditioned stimuli. *Royal Society Open Science*, *4*, 160935. https://doi.org/10.1098/rsos.160935

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of Rydell et al. (2006). *Cognition and Emotion*, *32*(8), 1708–1727. https://doi.org/10.1080/02699931.2018.1429389

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional Versus Dual-Process Accounts of Evaluative Conditioning: I. The Effects of Co-Occurrence and Relational Information on Implicit and Explicit Evaluations. *Personality and Social Psychology Bulletin*, *43*(1), 17–32. https://doi.org/10.1177/0146167216673351

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology*, *33*(1), 36–67. https://doi.org/10.1016/0022-2496(89)90003-5

Hütter, M., & Sweldens, S. (2013). Implicit misattribution of evaluative responses: Contingency-unaware evaluative conditioning requires simultaneous stimulus presentations. *Journal of Experimental Psychology: General*, *142*(3), 638–643. https://doi.org/10.1037/a0029989

Hütter, M., Sweldens, S., Stahl, C., Unkelbach, C., & Klauer, K. C. (2012). Dissociating contingency awareness and conditioned attitudes: Evidence of contingency-unaware evaluative conditioning. *Journal of Experimental Psychology: General*, *141*(3), 539–557. https://doi.org/10.1037/a0026477

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*(1), 61–82. https://doi.org/10.3758/s13420-011-0046-2

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*(5), 933–948. https://doi.org/10.1037/a0014747

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Kattner, F. (2014). Reconsidering the (in)Sensitivity of evaluative conditioning to reinforcement density and CS contingency. *Learning and Motivation*, *45*, 15–29. https://doi.org/10.1016/j.lmot.2013.09.002

Kattner, F., & Green, C. S. (2015). Cue competition in evaluative conditioning as a function of the learning process. *Acta Psychologica*, *162*, 40–50. https://doi.org/10.1016/j.actpsy.2015.09.013

Kelly, M. A., Mewhort, D. J. K., & West, R. L. (2017). The memory tesseract: Mathematical

equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142–155. https://doi.org/10.1016/j.jmp.2016.10.006

Klauer, K. C. (2009). Spontaneous evaluations. In F. Strack & J. Förster (Eds.), *Social cognition: The basis of human interaction* (pp. 199–217). New York, NY: Psychology Press.

Klauer, K. C., & Teige-Mocigemba, S. (2007). Controllability and resource dependence in automatic evaluation. *Journal of Experimental Social Psychology*, *43*(4), 648–655. https://doi.org/10.1016/j.jesp.2006.06.003

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550617697177

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.

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Lipp, O. V., Mallan, K. M., Libera, M., & Tan, M. (2010). The effects of verbal instruction on affective and expectancy learning. *Behaviour Research and Therapy*, *48*(3), 203–209. https://doi.org/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*(3), 219–239. https://doi.org/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*(4), 431–440. https://doi.org/10.1037/0097-7403.32.4.431

Lovibond, P. F. (2004). Cognitive Processes in Extinction. *Learning & Memory*, *11*(5), 495–500. https://doi.org/10.1101/lm.79604

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26. https://doi.org/10.1037/0097-7403.28.1.3

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matute, H., Lipp, O. V., Vadillo, M. A., & Humphreys, M. S. (2011). Temporal contexts: Filling the gap between episodic memory and associative learning. *Journal of Experimental Psychology: General*, *140*(4), 660–673. https://doi.org/10.1037/a0023862

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*(4), 714–725. https://doi.org/10.1037/0278-7393.28.4.714

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*(3), 363–386. https://doi.org/10.1037/0033-2909.117.3.363

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183–198. https://doi.org/10.1017/S0140525X09000855

Moran, T., & Bar-Anan, Y. (2013). The effect of objectvalence relations on automatic evaluation. *Cognition and Emotion*, *27*(4), 743–752. https://doi.org/10.1080/02699931.2012.732040

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Olson, M. A., & Fazio, R. H. (2001). Implicit Attitude Formation Through Classical Conditioning. *Psychological Science*, *12*(5), 413–417. https://doi.org/10.1111/1467-9280.00376

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*(1), 130–144. https://doi.org/10.1037/0278-7393.33.1.130

Raes, A. K., De Houwer, J., Verschuere, B., & De Raedt, R. (2011). Return of fear after retrospective inferences about the absence of an unconditioned stimulus during extinction. *Behaviour Research and Therapy*, *49*(3), 212–218. https://doi.org/10.1016/j.brat.2010.12.004

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Richter, J., & Gast, A. (2017). Distributed practice can boost evaluative conditioning by increasing memory for the stimulus pairs. *Acta Psychologica*, *179*, 1–13. https://doi.org/10.1016/j.actpsy.2017.06.007

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*. https://doi.org/10.1037/met0000061

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321. https://doi.org/10.3758/BF03210755

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). *Afex: Analysis of factorial experiments*. Retrieved from https://github.com/singmann/afex

Stahl, C., Haaf, J., & Corneille, O. (2016). Subliminal Evaluative Conditioning? Above-Chance CS Identification May Be Necessary and Insufficient for Attitude Learning. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000191

Stahl, C., & Heycke, T. (2016). Evaluative Conditioning with Simultaneous and Sequential Pairings Under Incidental and Intentional Learning Conditions. *Social Cognition*, *34*(5), 382–412. https://doi.org/10.1521/soco.2016.34.5.382

Stahl, C., Unkelbach, C., & Corneille, O. (2009). On the respective contributions of awareness of unconditioned stimulus valence and unconditioned stimulus identity in attitude formation through evaluative conditioning. *Journal of Personality and Social Psychology*, *97*(3), 404–420. https://doi.org/10.1037/a0016196

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Sweldens, S., Corneille, O., & Yzerbyt, V. (2014). The Role of Awareness in Attitude Formation Through Evaluative Conditioning. *Personality and Social Psychology Review*, *18*(2), 187–209. https://doi.org/10.1177/1088868314527832

Teige-Mocigemba, S., & Klauer, K. C. (2008). "Automatic" evaluation? Strategic effects on affective priming. *Journal of Experimental Social Psychology*, *44*(5), 1414–1417. https://doi.org/10.1016/j.jesp.2008.04.004

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*(1), 71–79. https://doi.org/10.1037/0097-7403.32.1.71

Vansteenwegen, D., Hermans, D., Vervliet, B., Francken, G., Beckers, T., Baeyens, F., & Eelen, P. (2005). Return of fear in a human differential conditioning paradigm caused by a return to the original acquistion context. *Behaviour Research and Therapy*, *43*(3), 323–336. https://doi.org/10.1016/j.brat.2004.01.001

Vervliet, B., Vansteenwegen, D., Baeyens, F., Hermans, D., & Eelen, P. (2005). Return of fear in a human differential conditioning paradigm caused by a stimulus change after extinction. *Behaviour Research and Therapy*, *43*(3), 357–371. https://doi.org/10.1016/j.brat.2004.02.005

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wellek, S. (2002). *Testing Statistical Hypotheses of Equivalence* (1 edition). Boca Raton, Fla: Chapman and Hall/CRC.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A Model of Dual Attitudes. *Psychological Review*, *107*(1), 101. https://doi.org/10.1037/0033-295X.107.1.101

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zanon, R., De Houwer, J., & Gast, A. (2012). Context effects in evaluative conditioning of implicit evaluations. *Learning and Motivation*, *43*(3), 155–165. https://doi.org/10.1016/j.lmot.2012.02.003

Appendix A
MINERVA 2 simulation method

We assumed that trials were encoded as combinations of stimulus and context features. CSs and USs consisted of 10 unique features coded as 1. Hence, for simplicity we assumed that all stimuli were unrelated to each other. For USs 10 additional features coded stimulus valence with 1 for positive, -1 for negative, and 0 for neutral or no valence. The contexts were represented by 10 common and 20 distinguishing features. For the first context, 10 distinguishing features were coded as 1, indicating the presence of some contextual features, and the remaining 10 features were coded as -1, indicating the absence of other contextual features. The coding was reversed for the second context. The context for end-of-study pleasantness ratings was represented by the 10 common and 40 unique features, all coded as 1—the distinguishing context features of the learning procedure were coded as 0. Thus, we assumed participants would experience the end-of-study rating procedure as markedly different from the learning procedure.

We further assumed that participants' memory initially contained information unrelated to the experiment. This assumption was implemented by starting with a memory containing 100 episodes where each feature was randomly coded as -1, 0, or 1. Each CS-US pairing was appended to the memory as a new trace and features were correctly encoded into memory with a probability of $p = 0.60$ or as 0 otherwise. We simulated 10 trials for each context (i.e. acquisition and counterconditioning or extinction). Each simulation was repeated 30 times.

To predict US expectancy and CS pleasantness ratings, we reasoned that the CS in question and the current context act as cues to recall previous pairings with USs. Hence, we used the CS and context to probe memory and computed the normalized memory echo, in which features range from -1 to 1. The normalized memory echo represents the recalled information—a mixture of all learning episodes involving the CS. We then determined the valence of the recalled content by averaging across the recalled valence-coding features. If the recalled content was positive we predicted an expectation of a positive US and a positive CS evaluation. Thus, we predicted US expectancy and CS pleasantness ratings based on the same information. This approach is essentially equivalent to predicting US expectancy from orthogonal category-specific features (e.g., human, animal, or object features).

Appendix B
CS-US pairing memory

Here we report the analysis of participants' US category and US identity recognition responses.

Experiment 1

We analyzed US category and identity recognition responses using 2 (*US valence order*: US+ US− vs. US− US+) × 2 (*Context*: First vs. Second) repeated-measures

ANOVAs.

Overall, US category recognition was quite accurate. We observed a small recency effect, that is, US category recognition was somewhat better for the second ($M = .87$, SD = .19) than for the first context ($M = .78$, SD = .22), $F(1, 36) = 10.44$, $MSE = 0.03$, $p = .003$, $\hat{\eta}^2_G = .051$, $\text{BF}_{10} = 76.76$. We found no noteworthy evidence for any other effects of our experimental manipulations, all $p \geq .245$, all $\text{BF}_{01} \geq 2.62$.

A one-way repeated-measures ANOVA of end-of-study pleasantness ratings of US categories indicated that participants remembered the valence of the US categories, $F(1.85, 66.57) = 115.47$, $MSE = 4.48$, $p < .001$, $\hat{\eta}^2_G = .710$, $\text{BF}_{10} = 3.71 \times 10^{27}$. Without any exemplars available, participants rated the animal category as more pleasant than the object category, $\Delta M = 2.81$, 95% CI [1.94, 3.68], $t(36) = 6.54$, $p < .001$, $\text{BF}_{10} = 5.01 \times 10^{6}$, and the human category as less pleasant than the object category, $\Delta M = -4.32$, 95% CI [−5.23, −3.42], $t(36) = -9.67$, $p < .001$, $\text{BF}_{10} = 2.26 \times 10^{14}$. Thus, recognition memory for US categories may be indicative of participants' US valence memory. Note, however, that participants rated US categories after the US identity recognition assessment during which we presented arrays containing all exemplars from each US category.

Recognition accuracy for the specific USs that had been paired with CSs followed a similar pattern. Overall, US identity recognition was quite accurate in both the first ($M = .73$, SD = .25) and the second context, $M = .82$, SD = .23. However, the observed recency effect in US identity recognition appeared to be largely due to CSs that had first been paired with positive and then with negative USs, $F(1, 36) = 9.48$, $MSE = 0.03$, $p = .004$, $\hat{\eta}^2_G = .029$, $\text{BF}_{10} = 26.39$. Participants were less accurate to recognize the positive USs that had been paired with CSs in the first learning phase than the corresponding negative USs from the first phase, $\Delta M = 0.17$, 95% CI [0.08, 0.26], $t(36) = 4.33$, $p < .001$ (adjusted for two comparisons), $\text{BF}_{10} = 219.60$. There was some evidence, however, that there was no recency effect for CSs that had first been paired with negative USs and later with positive USs, $\Delta M = 0.01$, 90% CI [−0.06, 0.07], $t(36) = -1.93$, $p = .061$ (equivalence test adjusted for three comparisons), $\text{BF}_{01} = 5.45$.

The memory-based judgment perspective assumes that EC requires memory of CS and US valence. We tested whether the observed changes in CS pleasantness across contexts was contingent on memory for CS-US pairs. Due to the overall high memory accuracy only small subsamples were available to test our hypotheses. Nonetheless, we found some evidence that the observed EC effects were contingent on memory for US categories, $F(1, 6) = 7.67$, $MSE = 6.52$, $p = .032$, $\hat{\eta}^2_G = .113$, $\text{BF}_{10} = 5.68$. This finding also corroborates that US category recognition is indicative of US valence memory. Our analyses regarding the role of memory for US identity were inconclusive, $F(1, 11) = 2.31$, $MSE = 5.29$, $p = .157$, $\hat{\eta}^2_G = .008$, $\text{BF}_{01} = 1.60$.

Experiment 2

Confirmatory results. We analyzed US category recognition accuracy using 2 (*Valence*: Positive vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) × 2 (*Context*: First vs. Second) repeated-measures ANOVA. As in Experiment 1, US category recognition was quite accurate. However, participants better remembered that no US had been presented (M = .92, SD = .18 and M = .93, SD = .20 for acquisition and extinction, respectively) than the correct US category when a CS had been paired with a US, M = .80, SD = .31 and M = .80, SD = .29 for acquisition and extinction, respectively, $BF_{10} = 1.66 \times 10^{12}$. Beyond the recognition advantage for US absence, we found evidence indicating that recognition performance was comparable between the learning procedures, $BF_{01} = 7.69$. We found no noteworthy evidence for any other effects of our experimental manipulations, all $BF_{10} \leq 2.21$.

We analyzed US identity recognition accuracy using 2 (*Valence*: Positive vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) repeated-measures ANOVA. US identity recognition, too, was quite accurate in both acquisition (M = .87, SD = .26) and extinction procedures, M = .85, SD = .27. We found no noteworthy evidence for any effects of our experimental manipulations, all $BF_{10} \leq 1.39$.

Exploratory results. As in Experiment 1, a one-way repeated-measures ANOVA of end-of-study pleasantness ratings of US categories indicated that participants remembered the valence of US categories, $BF_{10} = 5.96 \times 10^{51}$. Without any exemplars available, participants rated the animal category as more pleasant than the object category, $BF_{10} = 5.88 \times 10^{14}$, and the human category as less pleasant than the object category, $BF_{10} = 1.63 \times 10^{29}$. Thus, recognition memory for US categories may be indicative of participants' US valence memory.

Memory for CS-US pairings was too accurate to test whether the observed differences in EC effects across referenced contexts was contingent on memory for CS-US pairs.

Experiment 3

Confirmatory analyses. We analyzed US category recognition accuracy using 2 (*Valence*: Positive vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) × 2 (*Context*: First vs. Second) × 2 (*DV order*: Pleasantness first vs. Expectancy first) ANOVA with repeated-measures on the first three factors. Again, US category recognition was quite accurate. We found that the effect of context on US category memory differed between learning procedure, $BF_{10} = 1.29 \times 10^{14}$. Unlike in Experiment 2, we found evidence indicating that the recognition advantage for US absence was dependent on the learning procedure, $BF_{10} = 250.44$. Participants best remembered that a US was absent in the acquisition procedure (M = .89, SD = .26); however, memory for US absence in the extinction procedure (M = .78, SD = .34) was comparable to the memory for the correct category when a CS had been paired with a US (M = .76, SD = .31, and M = .74, SD = .32 for acquisition and extinction, respectively). These results were not affected by DV order, $BF_{01} = 8.13$; we found evidence that there were no other effects of our experimental manipulations, all $BF_{01} \geq 6.19$.

We analyzed US identity recognition accuracy using 2 (*Valence*: Positive vs. Negative) × 2 (*Learning procedure*: Acquisition vs. Extinction) × 2 (*DV order*: Pleasantness first vs. Expectancy first) ANOVA with repeated-measures on the first two factors. US identity recognition, too, was quite accurate in both acquisition ($M = .85$, $SD = .28$) and extinction procedure ($M = .85$, $SD = .28$). We found weak evidence suggesting that memory for negative USs ($M = .87$, $SD = .26$) was better than for positive USs ($M = .83$, $SD = .29$, $BF_{10} = 3.46$) but there was no noteworthy evidence indicating that any other experimental manipulation affected US identity recognition, all $BF_{10} \leq 1.80$.

Exploratory analyses. As in the previous experiments, a one-way repeated-measures ANOVA of end-of-study pleasantness ratings of US categories indicated that participants remembered the valence of the US categories, $BF_{10} = 2.11 \times 10^{150}$. Participants remembered the animal category as more pleasant than the object category, $BF_{10} = 4.04 \times 10^{34}$, and human category as less pleasant than object category, $BF_{10} = 1.15 \times 10^{79}$. Thus, recognition memory for US categories may be indicative of participants' US valence memory.

Memory for CS-US pairings again was too accurate to test whether the observed differences in EC effects across referenced contexts was contingent on memory for CS-US pairs.

Appendix C
Normative IAPS ratings for USs

Table C1

Identifiers of IAPS pictures used in CS-US and as filler USs with mean normative pleasure and arousal ratings (standard deviations in parentheses).

	CS-US pairs			
	Positive	Neutral	Negative	US-US pairs
	1610	7000	2750	9280
	1604	7035	2312	5970
	1620	7002	3300	5611
	1600	7009	2900.1	5250
	1750	7004	2276	5660
	1500	7233	2753	5870
	1460	7090	2110	5720
	1721	7080	9041	5780
	1540	7006	9331	9000
	1440	7175	2399	
	1463	7705	2100	
	1590	7025	2455	
Pleasure	7.56 (0.44)	4.97 (0.17)	3.12 (0.49)	5.61 (1.94)
Arousal	4.15 (0.57)	2.53 (0.40)	4.36 (0.28)	3.95 (0.74)

Note. IAPS = International Affective Picture System (Lang, Bradley, & Cuthbert, 2008), CS = Conditioned stimulus, US = Unconditioned stimulus

Table C2

Weighted means (and standard deviations) of normative US pleasantness in Experiment 1.

Valence	Pleasure	Arousal
Acquisition		
Positive	7.55 (0.42)	4.23 (0.44)
Negative	2.81 (0.48)	4.54 (0.36)
Counterconditioning		
Positive	7.54 (0.42)	4.20 (0.45)
Negative	2.82 (0.44)	4.52 (0.37)

Note. Normative ratings from Lang et al. (2008). US = Unconditioned stimulus

References

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.