Evaluative conditioning: missing, presumed dead

Andy P. Field, Kristy R.R. Lascelles*, Kathryn J. Lester, Chris Askew** and Graham C. L. Davey

Although research into evaluative conditioning (EC) has progressed considerably over the past 30 years, there have been some doubts about the strength and reliability of conditioning effects – especially when visual stimuli are used. Partly this has been due to well-documented methodological debates and empirical evidence of alternative causes of apparent learning (Field & Davey, 1999). The seeds of doubt have undoubtedly been nurtured, also by the numerous informal reports of researchers failing to obtain EC effects – even when replicating previously successful paradigms. This has led some to talk of boundary conditions that could enhance or eliminate the effects in EC experiments (De Houwer, Baeyens, Vansteenwegen & Eelen, 2000). This paper summarises 12 experiments that have explored this elusive phenomenon with mixed results (even when replicating past procedures). These various experiments are used to attempt to draw conclusions about what boundary conditions might exist for evaluative conditioning. (*Netherlands Journal of Psychology*, 64, 46-64.)

Keywords: valuative conditioning

Preferences have a pervasive impact on a wide range of behaviours, and most preferences are learned, not innate. Associative learning has become the dominant theory of preference learning; it is assumed that stimuli become liked or disliked through association with stimuli that we already like or dislike. This form of learning is known as evaluative conditioning (EC) and has been demonstrated using a variety of stimuli including visual images, words, tastes, olfactory stimuli, and auditory stimuli (see De Houwer, Thomas & Baeyens, 2001; Field, 2005; Havermans & Jansen, 2007 for reviews). Procedurally EC is a form of Pavlovian conditioning in which a conditioned stimulus, CS (in this context an affectively neutral stimulus), is paired with either a liked or disliked unconditioned stimulus (US), resulting in the CS evoking a response congruent with the US with which it was paired. Although procedurally the same as Pavlovian learning, EC is intriguing because it appears to

Department of Psychology, University of Sussex, Brighton
* Now at the Department of Psychology, University of Northampton
** Now at the Faculty of Arts and Social Sciences, Kingston University, Kingston upon Thames
Correspondence to Andy P. Field, Department of Psychology, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH, UK, e-mail: andyf@sussex.ac.uk.
Received 30 January 2008, revision accepted 2 April 2008.

have several unusual characteristics: at face value, it can occur without participants possessing awareness of the CS-US contingencies (see Field, 2000b for a review), responses acquired through EC are resistant to extinction (Baeyens, Crombez, van den Bergh & Eelen, 1988; Baeyens, Eelen & Crombez, 1995; Diaz, Ruiz & Baeyens, 2005), it can occur in few trials (Baeyens, Eelen, Crombez & Vandenbergh, 1992) and does not depend on CS-US contingency (Baeyens, Hermans & Eelen, 1993).

In terms of how EC works, there have been several models proposed (see De Houwer et al., 2001; Field, 2005; Havermans & Jansen, 2007 for reviews). To summarise, Baeyens and his colleagues have suggested that EC is based on referential connections between CS and USs, which, unlike Pavlovian learning, are not accompanied by expectancies that the CS predicts the occurrence of the US (Baeyens & De Houwer, 1995; Baeyens, et al., 1995). Martin and Levey (1994) have suggested that when the US is first experienced, the characteristics of the CS and US are laid down cognitively in a stimulus complex. This complex does not discriminate between the CS and US as separate stimuli, but places them in a 'holistic representation in which [the] CS, the UC and, probably, the evaluative response itself are fused or integrated' (Martin & Levey, 1994, p. 302). This model remains largely untested because it offers little in the way of testable mechanisms through which this 'holistic representation' is formed. Davey (1994) has suggested that EC is the result of conceptual learning and transfer. In essence, he argues that the EC procedure draws participants' attention to the features in the preceding neutral CS that are congruent with the liked or disliked US. For example, if a neutral CS is paired with a liked US, then the conceptual features of the CS that are liked become salient resulting in it being re-categorised as a liked stimulus. Generally speaking, neutral CSs are re-categorised congruent to the USs with which they are paired because the congruent conceptual features present in the CS become more salient than the incongruent conceptual features. A similar model has been proposed by Havermans and Jansen (2007) who apply Pearce's model of stimulus generalisation (Pearce, 1987). In this model, when a neutral CS is paired with a US the co-activation of their representations leads to a configural representation being formed. Subsequent experience of the CS evokes a representation of the US both directly and via the configural representation. In essence, a hedonic shift to the CS occurs through the generalisation of affect from the US to the CS. Like Davey's model, similarity between the CS and US is key (in Pearce's model the degree of stimulus generalisation is determined by the number of features that two stimuli share): the strength of responses acquired through EC will depend on the similarity between the CS and US.

Despite these interesting theoretical ideas there has been relatively little research that directly tests these models. In part, theoretical research has been hampered by an inconvenient undercurrent of findings that question both whether EC is an associative phenomenon and its very existence. For example, conditioning-type effects have also been shown in between-participant control conditions leading some to suggest that some of the available evidence may not reflect associative learning at all, but merely experimental artefacts (Field & Davey, 1997, 1998, 1999; Shanks & Dickinson, 1990). Furthermore, failures to obtain EC both in the laboratory and real-world settings have been reported (Field & Davey, 1999; Rozin, Wrzesniewski & Byrnes, 1998). Based on personal communications, Field (1997) reports 19 unpublished failures to obtain EC from a variety of laboratories (Field & Davey, Baeyens and colleagues, and Rozin and colleagues). As such, the extent to which EC is a reliable and robust phenomenon has been questioned, leading some to talk of boundary conditions that may determine whether EC is successful or not (De Houwer, Baeyens, Vansteenwegen & Eelen, 2000). For theories of EC to advance, these boundary conditions need to be better understood and theoretical models need to be tested. This paper brings together a series of 12 experiments, done over a period of time, which have manipulated various parametric features of the EC procedure to see what factors, if any, affect the strength of EC. The experiments fall into three categories, which reflect broadly the rationale for doing them: replications of published experiments with visual stimuli incorporating manipulations of basic parameters, replications of experiments with taste stimuli, and attempts to manipulate a theoretically important variable (the conceptual similarity between the CS and US).

Manipulating basic parameters using visual stimuli

Successful EC may depend upon certain boundary conditions being met (De Houwer, Baeyens & Field, 2005; De Houwer et al., 2000; De Houwer et al., 2001; Rozin et al., 1998), which may be determined by the basic parameters of the conditioning procedure. Particular parameters of EC procedures could be predicted to reduce successful associative learning, namely the tendency to use a larger number of different CS-US pairings (making the procedure more comparable to a covariation assessment procedure than a simple contiguous associative learning procedure), use of trace conditioning procedures which involve a significant delay between CS offset and US onset, and use of stimulus durations which may be too brief for the participant to process relatively complex visual stimulus. Experiments 1 to 6 explore some of these parametric issues. Experiment 1 is a close replication of Baeyens et al. (1988), Experiments 2 to 4 examine the effect of manipulating the timing parameters of CS and US presentations, Experiments 5 and 6 reduce the number of conditioning trials from ten to one to see if this enables simple one-trial learning of nine different CS-US pairings. Experiment 5 increases CS and US presentation times to 20 seconds in an attempt to provide participants with a better opportunity to fully process the information contained in relatively complex visual stimuli. Finally, Experiment 6 still looked at one-trial learning but decreased the number of CS-US pairs from nine to four to reduce attentional load.

Experiment 1

Experiment 1 closely replicated Baeyens et al.'s (1988) study examining resistance to extinction of evaluative responses. Differential evaluative responding was expected between CSs paired with liked and disliked USs with these acquired responses remaining after the CSs have been presented several times without reinforcement.

Method

Participants. Fifteen participants completed the experiment (age $M = 27.2$ years, $SD = 7.44$) and were paid a small sum for participation.

Stimuli. Seventy colour images of human faces (48 females, 22 males) developed for this experiment were used. Each image consisted of a passport photo scanned into a computer and edited to remove all of the background colour – leaving just the head and shoulders against a white background.

Apparatus. The experiment was run using a custom-written computer programme by the first author (Field, Matthias, Siddens-Corby & Ives, 1997).

Design. A 3 (US type: Neutral-Like (N-L), Neutral-Dislike (N-D), Neutral-Neutral (N-N)) × 3 (phase: baseline, postconditioning, postextinction) repeated measures factorial design was used.

Procedure. The procedure used was similar to Baeyens et al. (1988), aside from the procedure being computer automated.

Stage 1: Baseline assessment. Participants rated each of the randomly presented 70 faces along a scale ranging from –100 (dislike) through 0 (neutral) to +100 (like), in intervals of 5. The three faces with the highest ratings and the three faces with the lowest ratings were selected to be used as USs. Twelve neutral images were selected, which were defined as pictures with a rating of zero. Where there were an insufficient number of pictures rated exactly at zero, the computer selected pictures with the next lowest ratings between ±10. In the unlikely event that there were still not enough neutrally rated pictures, the computer selected pictures with the lowest ratings between ±20. If there were still not the required number of neutral faces, the experiment was terminated, which occurred on only two occasions. Nine stimulus pairs were constructed on an entirely random basis: 3 × Neutral-Like (N-L), 3 × Neutral-Dislike (N-D), and 3 × Neutral-Neutral (N-N).

Stage 2: Conditioning. The nine stimulus pairings were presented in a semi-randomised order (i.e. randomised but with the constraint that the same pair could not appear on more than two consecutive occasions). Each pair was presented ten times. Each stimulus appeared on the screen for 1 second, the interval between stimuli in a pair (the trace interval) was 4 seconds and the interval between the offset of one pair and the onset of the next pair (the inter-trial interval, or ITI) was 8 seconds.

Stage 3: Postconditioning assessment. The CSs and USs from the conditioning stage were presented again for rating. To eliminate the chance possibility of CS-US pairs appearing consecutively at this stage, the stimuli were presented in blocks (all of the CSs followed by all of the USs or vice versa). Stimuli within each block were presented randomly. At the end of this stage there was a 30 second gap.

Stage 4: Extinction. The nine CSs were presented alone, in semi-randomised order (random but with the restriction that a CS could not appear consecutively more than twice), without contingent presentations of their respective USs. The stimulus duration was 1 second, the time between the offset of one CS and the onset of the next was set at 8 seconds, and each CS was presented ten times.

Stage 5: Postextinction assessment. This was identical to stage 3.

Results

Data from one participant were excluded as this participant admitted to deliberately giving maximum negative scores in the postconditioning and postextinction assessments. The mean evaluative ratings of the USs were 53.10 (N-L pairs, $SE = 6.25$), –45.36 (N-D pairs, $SE = 5.45$) and 0.83 (N-N pairs, $SE = 0.49$). A one-way repeated measures ANOVA showed that these differences were significant, $F(2, 26) = 86.36$, $p < 0.001$: liked USs ($F(1, 13) = 75.15$, $p < 0.001$), and disliked USs ($F(1, 12) = 71.66$, $p < 0.001$) were rated significantly different from neutral ones.

Table 1 reports the CS ratings at baseline, after conditioning and after extinction, for each of the pair types. There was virtually no shift in CSs paired with liked USs and contrary to Baeyens' findings, CSs paired with disliked and neutral USs shifted in a positive direction.

Table 1 **Mean evaluative responses for CSs paired with liked or disliked US across the first six experiments (*SE* is in parenthesis).**

		Experimental			*BSB control/unaware*		
Experiment	*Pair type*	*Baseline*	*Post-conditioning*	*Post-extinction*	*Baseline*	*Post-conditioning*	*Post-extinction*
Experiment 1	N-L	0.36 (0.58)	0.12 (3.69)	−0.24 (3.61)	–	–	–
	N-D	0.24 (0.46)	6.19 (5.68)	−0.12 (0.48)	–	–	–
	N-N	10.12 (5.50)	9.52 (5.67)	5.95 (5.69)	–	–	–
Experiment 2	N-L	1.78 (0.65)	12.78 (3.99)	10.23 (4.02)	−0.33 (0.65)	6.89 (3.99)	7.22 (4.02)
	N-D	2.00 (0.63)	9.44 (2.50)	7.44 (2.78)	0.11 (0.63)	7.89 (2.50)	7.78 (2.78)
	N-N	−0.44 (0.51)	6.22 (3.56)	6.67 (3.68)	0.00 (0.51)	8.11 (3.56)	8.11 (3.68)
Experiment 3	N-L	1.73 (0.96)	16.73 (6.29)	–	–	–	–
	N-D	0.58 (0.58)	13.46 (4.72)	–	–	–	–
	N-N	0.39 (0.93)	4.42 (6.23)	–	–	–	–
Experiment 4	N-L	4.17 (2.46)	0.18 (2.23)	–	–	–	–
	N-D	5.54 (2.08)	0.83 (1.28)				
	N-N	3.87 (2.16)	3.51 (2.16)	–	–	–	
Experiment 5*	N-L	1.28 (1.68)	0.58 (1.33)	–	5.56 (4.75)	0.56 (0.56)	–
	N-D	4.52 (1.35)	3.41 (1.34)	–	0.00 (0.00)	3.89 (3.89)	–
	N-N	2.95 (1.31)	1.91 (1.32)	–	8.89 (4.75)	5.00 (3.47)	–

			Experimental			BSB control/unaware		
Experiment	Pair type		Baseline	Post-conditioning	Post-extinction	Baseline	Post-conditioning	Post-extinction
Experiment 6	N-L		6.22 (2.27)	6.19 (1.79)	–	-1.81 (2.93)	2.82 (2.92)	–
	N-D		2.16 (2.08)	3.45 (2.19)	–	4.49 (2.71)	1.16 (3.20)	–

Experiments with asterisks are split by contingency awareness, not by experimental and control group.

A two-way 3 (US type: N-L, N-D, N-N) × 3 (phase: baseline, postconditioning, postextinction) repeated measures ANOVA was conducted on the CS ratings. The US type, $F(1.32, 17.10) = 2.26$ and phase, $F(2, 26) < 1$ main effects were nonsignificant as was the US type × phase interaction, $F(1.60, 20.73) < 1$. In short, no evaluative conditioning was observed. Calculations on the interaction term comparing changes in ratings across conditioning between the N-L and N-D pairs revealed an effect size of $r = 0.26$, $(F\ 1, 13) < 1$, which would require a sample size of 50 to detect. This suggests that the failure to obtain conditioning is not simply because of a smaller sample size than used by Baeyens et al. (1988).

Experiment 2

Experiment 2 examined whether resistance to extinction could be elicited compared with a non-paired control condition – the Block/SubBlock (BSB) Control (Field, 1997) and investigated if reduction of the trace interval would facilitate conditioning by helping participants to identify distinct pairings. The use of a BSB control eliminates all CS-UC associations, holds presentational factors constant and controls for effects resulting from stimulus selection procedures and as such is useful in allowing conclusions to be drawn about the associative nature of any conditioning effects observed. In Baeyens et al.'s original study, images were projected using a slide projector; therefore, there would have been subtle audio cues at the onset and offset of a stimulus, which may help participants to detect the relative temporal difference between the trace interval and the ITI enabling them to perceive pairs of stimuli rather than a random display of single stimuli. The transition of images in our replication was silent. In this second replication, therefore, we reduced the trace interval to try to make CS-US pairings more obvious (relative to the ITI).

Method

Participants. Fifteen participants were assigned to an experimental group (age $M = 24.60$, $SD = 9.85$ years) and 16 participants were assigned to the BSB control (age $M = 28.56$, $SD = 6.43$). All participants were paid a small sum for their participation. Data from one participant was excluded as they did not rate enough pictures as neutral to construct pairings.

Procedure. The experimental procedure was identical to Experiment 1 except the trace interval was reduced from 4 seconds to 200 milliseconds.

The BSB control: CSs and USs were selected and paired together randomly. However, during conditioning the CSs and USs were not presented in a contiguous or contingent pattern. Instead, each stimulus was presented with itself, as a pairing, five times using the same parameters as the experimental condition. Thus, a stimulus appeared for 1 second, followed by a blank screen for 200 ms, followed by the same stimulus presented for 1 second, followed by a blank screen for 8 seconds, and so on until that stimulus had been presented ten times. This set of self-presentations is a block of pairings. Nine CS blocks and nine US blocks were presented.

Half of the participants saw the nine CS blocks presented in random order followed by the nine US blocks in random order, and half received the opposite order. Separating the CS blocks from the US blocks ensured that participants never saw a CS appearing contingently with a US. Randomising the presentation order of the blocks eliminated the possibility that participants could detect the US that corresponded to a CS.

Design. A 3 (US type: N-L, N-D, N-N) × 3 (phase: baseline, postconditioning, or postextinction) × 2 (group: paired or BSB control) mixed factorial design was used with US type and phase variables as repeated measures and group as a between-group factor.

Results
The mean evaluative ratings of the USs were 53.56 (N-L pairs, $SE = 5.21$), –59.00 (N-D pairs, $SE = 5.46$) and 0.56 (N-N pairs, $SE = 0.85$) in the experimental group and 53.78 (N-L pairs, $SE = 5.21$), –41.33 (N-D pairs, $SE = 5.46$) and –0.22 (N-N pairs, $SE = 0.85$) in the BSB control group. A two-way US type (N-L, N-D, N-N) × group (experimental or BSB control) ANOVA revealed that these differences were significant, $F(2, 56) = 227.06$, $p < 0.001$: liked USs ($F(1, 28) = 181.00$, $p < 0.001$), and disliked USs ($F(1, 28) = 172.54$, $p < 0.001$) were both rated significantly different from neutral ones. There was no significant group × US type interaction, $F(2, 56) = 2.27$.

A three-way 3 (US type: N-L, N-D, N-N) × 3 (phase: baseline, postconditioning, or postextinction) × 2 (group: paired or BSB control) ANOVA revealed nonsignificant main effects of group, $F(1, 28) < 1$, and US type, $F(2, 56) < 1$ and nonsignificant interactions between group × US type, $F(2, 56) < 1$, phase × US type, $F(1.88, 52.51) < 1$, group × phase, $F(1.02, 28.61) < 1$, and group × US type × phase interaction, $F(1.88, 52.51) < 1$. The only significant result was the main effect of phase, $F(1.02, 28.61) = 15.01$, $p < 0.001$. Contrasts showed that ratings after the conditioning stage differed significantly from those at baseline, $F(1, 28) = 17.16$, $p < 0.001$ and that ratings at extinction differed significantly from ratings after conditioning, $F(1, 28) = 6.30$, $p < 0.05$. Calculations on the interaction term comparing changes in ratings across conditioning between the N-L and N-D pairs in the experimental compared with the control group revealed an effect size of $r = 0.12$, $F(1, 28) < 1$, which would require a sample size of 150 to detect. The failure to obtain conditioning is not simply because of a smaller sample size than used by Baeyens et al., (1988).

In summary, although CS ratings shifted significantly between baseline and postconditioning, this was simply the result of repeated exposure to the stimuli as the type of US with which a CS was paired did not influence its subsequent ratings (table 1). Postexperimental interviews revealed that reducing the trace interval did make the pairings more apparent, but this awareness was insufficient to generate conditioned responding.

Experiment 3

Experiment 3 attempted to replicate Baeyens, Eelen & van den Bergh (1990). In addition to replicating the contingency awareness procedures that Baeyens et al. (1990) used, a recognition measure was included as a more sensitive test of awareness (Field, 2000b). This experiment explored whether previous failures to elicit conditioning effects in Experiments 1 and 2 can be explained by a lack of contingency awareness. By reducing the number of pairings used during conditioning, and therefore reducing the complexity of the conditioning task, it was hoped that levels of awareness could be enhanced.

Method
Participants. Thirteen participants completed the experiment and were paid a small sum for their participation (age $M = 21.38$, $SD = 2.26$ years).

Procedure. An identical procedure to Experiment 1 was used except the trace interval was reduced from 4 seconds to 200 milliseconds, there were no extinction trials (instead, 2 awareness measures were used) and only six CS-US pairs were used rather than nine: 2 × Neutral-Like (N-L), 2 × Neutral-Dislike (N-D), and 2 × Neutral-Neutral (N-N). These pairs were constructed randomly.

Two measures of contingency awareness were administered in a counterbalanced order following postconditioning ratings. In the first measure taken from Baeyens et al. (1990a), each CS was presented and participants were asked to identify the US from the remaining CS and US pictures and then indicate the certainty of their choice (completely sure, rather sure, rather unsure, completely unsure). In a second recognition measure (Field, 2000) participants were expected only to differentiate actual CS-US pairings from decoy CS-US pairings. Twelve pairs of faces were presented, some of which were actual CS-US pairs (6 pairs 2 × N-L, 2 × N-D, 2 × N-N) from the conditioning stage, whereas others were pairs containing a genuine CS but a decoy US (six pairs). The decoy US faces were selected from the preconditioning stage to have the same affective value as the actual US used during conditioning; none of these decoy USs appeared during conditioning. Participants indicated for each pairing whether they definitely remembered seeing that pair during the presentation stage, whether they had a gut-feeling that they had seen the pair but did not remember it, or if they definitely knew that they had not seen that pair.

Results
The mean US ratings were 66.35 (N-L pairs, $SE = 4.35$), –66.15 (N-D pairs, $SE = 4.88$) and 0.19 (N-N pairs, $SE = 1.22$). A one-way repeated measures ANOVA showed that these differences were significant, $F(2, 24) = 267.58$, $p < 0.001$: liked USs ($F(1, 12) = 214.51$, $p < 0.001$), and disliked USs ($F(1, 12) = 179.17$, $p < 0.001$) were both rated significantly different from neutral ones.

There was a strong positive shift in CS ratings regardless of the type of US with which it was paired (table 1). A two-way US type (N-L, N-D, N-N) × time (pre- or post-conditioning related ANOVA showed that the main effect of time was significant, $F(1, 12) = 7.76$, $p < 0.05$, showing that the ratings of CSs significantly increased across conditioning. However, the US type main effect, $F(2, 24) = 1.55$, and the crucial US type × time interaction, $F(2, 24) = 1.45$, were nonsignificant. Calculations on the interaction term comparing changes in ratings across time in N-L compared with N-D pairs revealed an effect size of $r = 0.11$, which would require a sample size of 150 to detect. Therefore, the current results do not merely reflect the use of a small sample size. The data cannot be explained by awareness levels either, because no participants were completely unaware of the contingencies (the awareness data only for this experiment are reported by Field, 2000).

Experiment 4

Experiment 4 is essentially a replication of Experiment 3 except that as well as having a shorter inter-stimulus interval (ISI) (500 ms), the ITI was also reduced (from 8000 ms in Baeyens et al.'s (1990) to 4000 ms in this experiment). In addition, the typical EC paradigm, where CSs and USs are selected based on a particular participant's evaluations of the stimulus set, may be prone to an experimental artefact that creates conditioning-type effects that are not due to associative learning (Field & Davey, 1997, 1998, 1999). Shanks and Dickinson (1990) indicate that this CS-UC assignment procedure may interact with the effects of stimulus exposure to create the illusion of conditioning. They argue that true conditioning effects can be isolated only if the pairing of a particular CS with a particular UC is counterbalanced across subjects. If this criterion is not met then it is possible that any conditioning-like effects could be due to non-associative factors arising from the paradigm. This artefact can be eliminated by using a counterbalanced, discriminative conditioning paradigm as employed in Experiment 4 in which the same CSs and USs are used across participants, but with CSs allocated to USs using counterbalancing. This procedure relies on using USs that are uniformly regarded as positive or negative, and CSs that are uniformly regarded as neutral. This can be done either by collecting ratings of a series of pictures in a pilot study, or by using stimuli that have been standardised, e.g. International Affective Picture System, IAPS (Lang, Bradley & Cuthbert, 1997). The second aim of Experiment 4 was, therefore, to explore whether true evaluative conditioning effects can be established using a paradigm in which nonassociative artefacts which occur as a result of biases in the way in which stimuli are selected and paired together are eliminated.

Method

Participants. Twenty-eight paid participants completed the experiment.

Stimuli. Eighteen photographs were selected from the IAPS; 12 pictures (9 neutral CSs and 3 neutral USs) with neutral pleasure ratings, three pictures (3 liked USs) with high pleasure ratings and three pictures (3 disliked USs) with low pleasure ratings.

Procedure. The procedure was ostensibly the same as Experiment 1 except that during the conditioning stage each CS was presented for 1000 ms followed by an ISI of 500 ms, after which the US was presented for 1000 ms. The ITI was 4 seconds. Each pair was shown for a total of 10 trials in random order.

Results

The mean preconditioning evaluative ratings for the liked, neutral and disliked USs were $M = 61.07$ ($SE = 4.05$), $M = 2.50$ ($SE = 2.03$), and $M = -72.38$ ($SE = 4.74$) respectively. A one-way repeated measures ANOVA showed a significant main effect of US type, $F(1.31, 35.27) = 237.72$, $p < 0.001$, was obtained with significant differences between neutral and liked, $F(1, 27) = 156.36$, $p < 0.001$, and neutral and disliked, $F(1, 27) = 215.55$, $p < 0.001$, USs.

Table 1 shows the mean CS ratings during the baseline and postconditioning assessment phases. The CSs paired with disliked USs became more disliked and the CSs paired with neutral USs showed little change; however, the CSs paired with liked USs also became more disliked following conditioning. A 3 (US type: liked, neutral, disliked) × 2 (time: preconditioning, postconditioning) repeated measures ANOVA revealed no significant main effects of either US type, $F(2, 54) < 1$, or time, $F(1, 27) = 3.85$, or the crucial US type × time interaction, $F(1.49, 40.18) = 2.63$, *ns*.

Experiment 5

This experiment was the same as Experiment 4 except that the number of CS-US trials was reduced from ten to one whilst increasing the duration of the CS and US presentation times from 1000 ms to 20 seconds. The rationale for these changes was to make the CS-US pairings as obvious as possible, whilst also making the cognitive load of conditioning trials as minimal as possible.

Method

Participants. Seventy-two paid participants completed the experiment (age $M = 27.50$, $SD = 9.08$).

Stimuli. Twenty-seven pictures were selected from the IAPS (Lang et al. 1997). These were the nine CS-US pairs used in Experiment 1 and nine decoy USs (3 L, 3 N, 3 D) for the recognition awareness test.

Procedure. The procedure was identical to Experiment 4, except that all 27 stimuli were rated during the baseline and postconditioning assessment phases. The CS and US presentation times were increased to 20 seconds and the number of presentation trials was reduced to one. Following the postconditioning rating phase, participants completed the awareness measures described in Experiment 3.

Results
The mean baseline US ratings revealed that the liked USs ($M = 55.46$, $SE = 3.09$) were rated as more liked than the neutral USs ($M = 0.69$, $SE = 1.58$), which were rated as more liked than the disliked USs ($M = -76.44$, $SE = 2.80$). An ANOVA revealed a significant main effect of US type ($F(1.35, 96.13) = 537.76$, $p < 0.001$) where the liked USs were significantly more liked than the neutral USs ($F(1, 71) = 233.13$, $p < 0.001$) and the disliked USs were significantly more disliked than the neutral USs ($F(1, 71) = 678.72$, $p < 0.001$).

Table 1 shows the mean CS ratings during the baseline and postconditioning assessment phases for aware ($n = 3$) and not aware participants ($n = 69$). For the aware participants, the CSs paired with liked, disliked and neutral USs became more disliked following conditioning. For the not aware participants, the CSs paired with liked and neutral USs became more disliked following conditioning and the CSs paired with disliked USs became more liked following conditioning.

A 3 (US type: liked, neutral, disliked) x 2 (phase: preconditioning, postconditioning) x 2 (awareness: aware, not aware) mixed ANOVA conducted on the CS evaluative ratings revealed no significant main effects of either US type, $F(2, 142) = 1.73$, or time, $F(1, 71) = 1.39$, *ns*. No significant differential conditioning effects were obtained as indicated by the nonsignificant US type and time interaction, $F(2, 142) < 1$.

Experiment 6

The number of different CS-US pairs was decreased to four (2 × N-L, 2 × N-D) to decrease attentional load still further. Filler stimuli were also presented for rating in the preconditioning and postconditioning phases, but were not presented during the conditioning phase.

Method
Participants. One hundred and thirty-six paid participants were recruited (age $M = 22.74$, $SD = 5.30$) with 82 participants randomly assigned to the experimental condition and 54 assigned to the BSB unpaired control condition.

Stimuli. Twenty pictures were selected from those used in Experiment 2; 12 neutral stimuli (4 CSs, and 8 neutral filler stimuli), four liked stimuli (2 USs and 2 decoy USs in the recognition awareness test) and four disliked stimuli (2 USs and 2 decoy USs in the recognition awareness test). The stimuli formed four CS-US pairs (2 N-L and 2 N-D pairs) and these pairs were fully counterbalanced such that all CSs were paired with all USs across participants.

Procedure. The procedure in the experimental group was identical to Experiment 5, except that only four CS-US pairs were presented. For the BSB control, each CS-CS and US-US pair was presented for one trial, but the stimulus presentation time was halved to 10,000 ms, the ISI was halved to 100 ms and the ITI was halved to 4000 ms maintaining overall identical exposure times in the BSB control condition to the experimental condition.

Results
The mean baseline evaluative US ratings showed that for the experimental condition the liked USs ($M = 55.67$, $SE = 2.87$) were more liked than the disliked USs ($M = -81.13$, $SE = 2.78$) and for the BSB control group the liked USs ($M = 62.08$, $SE = 3.07$) were more liked than the disliked USs ($M = -83.80$, $SE = 3.00$). A 2 (US type: liked, disliked) x 2 (condition: experimental, BSB control) ANOVA showed that across both groups the liked USs were significantly more liked than the disliked USs, $F(1, 134) = 1877.56$, $p < 0.001$. No significant main effect of condition, $F(1, 134) < 1$, or significant interaction of US type and condition, $F(1, 134) = 1.94$, *ns* was obtained.

Table 1 shows that for the experimental group, the CSs paired with liked USs showed no change and the CSs paired with disliked USs became slightly more liked following conditioning. For the BSB group, the CSs paired with liked USs became more liked and the CSs paired with disliked USs became more disliked following conditioning. A 3 (US type: liked, neutral, disliked) × 2 (phase: preconditioning, postconditioning) x 2 (condition: experimental, BSB) mixed ANOVA conducted on the CS evaluative ratings revealed a significant interaction of US type, time and condition, $F(1, 34) = 4.71$, $p < 0.05$. Separate US type × time ANOVAs were conducted on the BSB group and the experimental group. The US type and time interaction was significant for the BSB control group, $F(1, 53) = 6.07$, $p < 0.05$), but not for the experimental group, $F(1,81) < 1$. These results show significant effects for the BSB group, but not the experimental group. As the BSB group received no CS-US pairings the effects observed in this group could not have been the result of associative learning. No other significant main effects or interactions were obtained.

Discussion
Six attempts using a visual paradigm failed to demonstrate any evidence of evaluative conditioning. The type of US with which a CS was paired did not influence its subsequent ratings in any of the experiments. The parametric manipulations across these six Experiments shed little light on whether the basic parameters of the conditioning procedure determine the appearance of EC. Reducing the trace interval and the number of pairings did not facilitate EC, and EC remained elusive even when (1) the ISI in the trace delay procedure was reduced from 4000 to

500 msecs, (2) the number of trials was reduced from ten to one to facilitate awareness of contingencies and enhance comparison between N-L and N-D pairings, (3) stimulus presentation times were increased from 1 to 20 seconds, and (4) the number of different CS-US pairs used was reduced from nine to four. Across these studies sample sizes were comparable to or larger than the original studies on which they were based, and when this was not the case, effect sizes were so small that they would require huge sample sizes to detect.

Replications using taste stimuli

Having failed to establish reliable EC effects using visual stimuli we decided to change modality and look at EC using taste stimuli. The rationale for this change was that effects using taste stimuli were stronger (based on personal communications with Frank Baeyens). Essentially, the reason for these studies was again to establish a reliable paradigm from which more interesting Experiments could be designed to test theoretical predictions from the various models. The two experiments reported are close replications of published taste EC experiments (Baeyens, Crombez, Hendrickx & Eelen, 1995; Baeyens, Eelen, van den Bergh & Crombez, 1990) using the original design (Experiment 7) and with the addition of a BSB control group to rule out experimental artefacts (Experiment 8).

Experiment 7

Method

Participants. Twenty-seven participants completed the experiment and were paid a small sum for their participation (age $M = 22.37$, $SD = 4.36$).

Stimuli. Two natural colourless food flavourings were diluted in mineral water at the following concentrations: 0.75 ml of apricot in 1.5 litres of water, and 1.09 ml of melon diluted in 1.5 litres of water. These concentrations were pretested to evoke neutral evaluative responses.

These drinks were coloured red, blue or green by adding either 0.5 ml of red (cochineal) food colouring to 1.5 litres of drink, 1 ml of blue food colouring to 1.5 litres of the drink, or 2 ml of green food colouring to 1.5 litres of the drink. All colourings were tasteless.

CS+ drinks were created by adding Tween 20 (polysorbate 20) at the concentration of 0.75 ml to 1.4 litres of the drink. Tween 20 is a harmless chemical that has a bitter aftertaste. CS– drinks had no chemical added and acted as a baseline against which to compare the effects of the Tween.

Apparatus. Each participant was seated at a table containing a series of transparent 30 ml cups. At the front, six plastic cups were labelled 1 to 6, which contained the drinks for the preconditioning ratings. A tray containing 24 cups labelled 1 to 24 contained the drinks for the conditioning stage. Behind this tray, a further nine drinks (labelled 1 to 9) were positioned in a line.

Procedure. The four stages of the experiment were as follows:

Preconditioning: Participants rated six drinks each containing 5 ml of liquid: three contained apricot flavouring and the remainder melon flavouring. For each flavour there were three different coloured drinks: red, green and blue. The drinks were randomly ordered. The experimenter cued the participant to pick up each drink and look at it for 5 seconds, then to place the liquid into their mouth and swirl it around for a further 5 seconds, then swallow. Participants rated how much they liked the drink on a visual-analogue scale ranging from –100 (dislike) through 0 (neutral) to 100 (like). Participants ate a small piece of bread to clear their pallet while resting for 30 seconds before tasting the next drink.

Conditioning: Participants consumed 24 5 ml samples of liquid: 12 melon flavoured (4 blue, 4 red and 4 green) and 12 apricot flavoured (4 of each colour). One of these flavours acted as a CS–, and these drinks merely contained the flavour and water, the other flavour acted as the CS + and had a negative US added (Tween 20), which produces a bitter aftertaste. For half of the participants the melon flavour was the CS+, and for the remainder apricot was the CS+. The order of the 24 drinks was randomised with the restriction that no flavour or colour could appear more than twice in succession. The drinks were sampled and rated as described for the preconditioning stage.

Postconditioning: Participants consumed nine 5 ml drinks. The first six were three apricot and three melon drinks (one of each colour), which were randomly ordered. The last three drinks contained CS+ drinks (one of each colour), that is, the CS+ flavour with Tween 20 added. All drinks were tasted and rated as in the preconditioning stage.

Results

The mean rating of the CS+ drinks was –51.73 ($SE = 6.89$), indicating that participants did find the US aversive. Table 2 shows the mean ratings of the drinks pre- and post-conditioning. Both CS+ and CS– drinks showed negative shifts in ratings across conditioning, although these shifts were more pronounced in CS+ drinks. A 2 (US type: CS+ or CS–) × 2 (phase: pre- or post-conditioning) repeated measures ANOVA was conducted on the data. (A more complex analysis was run that included colour of the drink and which flavour acted as the CS+ as factors. These factors did not significantly effect changes in CS ratings and so the simpler analysis is reported.)

There was no significant main effect of US type, $F(1, 26) < 1$, or phase, $F(1, 26) = 1.55$. The crucial US type × phase interaction was also nonsignificant and yielded a small effect size, $F(1, 26) = 1.10$, $r = 0.20$ indicating that shifts in CS ratings across conditioning did not depend on the type of US used.

Table 2 **Mean evaluative responses for CSs paired with liked or disliked US for experiments 7 to 12 (SE is in parenthesis).**

		*Experimental/aware**		*BSB control/unaware**	
Experiment	*Pair type*	*Baseline*	*Post-conditioning*	*Baseline*	*Post-conditioning*
Experiment 7	CS+	−9.14 (5.29)	−17.04 (6.65)		
	CS−	−14.32 (5.88)	−15.86 (6.75)		
Experiment 8	CS+	−3.96 (5.37)	−6.67 (5.34)	6.77 (5.37)	11.98 (5.34)
	CS−	−4.06 (6.10)	−10.73 (5.21)	−0.94 (6.10)	7.71 (5.21)
Experiment 9	N-L	2.21 (5.18)	−8.09 (4.69)	–	–
	N-D	−4.71 (6.46)	−18.68 (5.01)	–	–
Experiment 10 (Overall)	N-L (Relevant)	-5.31 (4.93)	0.31 (4.03)	−2.50 (4.93)	4.06 (4.03)
	N-D (Relevant)	4.38 (5.56)	−1.88 (2.50)	1.25 (5.56)	3.13 (2.50)
	N-L (Irrelevant)	−0.31 (5.83)	−4.38 (5.15)	−10.94 (5.83)	−12.19 (5.15)
	N-D (Irrelevant)	−6.25 (7.01)	−0.63 (5.80)	3.44 (7.01)	4.69 (5.80)
Experiment 10* (Awareness)	N-L (Relevant)	−12.50 (8.14)	−5.00 (3.42)	−1.00 (6.94)	3.50 (4.72)
	N-D (Relevant)	9.17 (7.57)	−3.33 (3.33)	1.50 (1.00)	−1.00 (1.25)
	N-L (Irrelevant)	−11.67 (9.46)	−5.00 (6.06)	6.50 (8.34)	−4.00 (5.86)
	N-D (Irrelevant)	−4.17 (13.69)	−4.17 (4.17)	−7.50 (10.57)	1.50 (8.40)

			Experimental/aware*		BSB control/unaware*	
Experiment		*Pair type*	*Baseline*	*Post-conditioning*	*Baseline*	*Post-conditioning*
Experiment 11		N-L (Relevant)	-6.33 (6.24)	3.00 (6.24)	3.75 (6.04)	1.88 (6.04)
		N-D (Relevant)	-9.00 (4.61)	-11.00 (4.68)	-4.38 (4.47)	-1.25 (4.53)
		N-L (Irrelevant)	-4.00 (5.66)	3.33 (4.48)	-1.88 (5.48)	-5.00 (4.34)
		N-D (Irrelevant)	-0.67 (5.26)	-14.00 (5.45)	-6.25 (5.09)	-4.38 (5.27)
Experiment 12		N-L (Relevant)	-8.38 (6.19)	-4.00 (3.75)	-1.47 (6.51)	2.00 (3.95)
		N-D (Relevant)	-4.38 (6.00)	-1.57 (4.13)	-6.21 (6.31)	-3.42 (4.34)
		N-L (Irrelevant)	-4.33 (2.76)	-7.00 (3.75)	0.00 (2.90)	-2.26 (3.93)
		N-D (Irrelevant)	-9.86 (4.90)	-3.29 (5.20)	1.90 (5.15)	2.95 (5.46)

Experiments with asterisks are split by contingency awareness, not by experimental and control group.

Experiment 8

Method

Participants. Sixty-four unpaid participants completed the experiment: 32 participants in the experimental group (age $M = 22.41$, $SD = 4.54$) and 32 participants in the BSB control (age $M = 21.16$, $SD = 3.92$).

Procedure. The procedure was identical to that of Experiment 7, aside from the inclusion of a BSB control (Field, 1997). As in the experimental group, BSB control participants consumed 24 5 ml drinks. Twelve of the drinks contained the two CS flavours (6 of each flavour, 2 of each colour). This was the CS block, and half of the participants tasted all of the melon drinks before the apricot ones, whereas the remainder tasted the drinks in the opposite order. The remaining 12 drinks contained water or water mixed with Tween 20 (6 of each drink, 2 of each colour). This was the US block, and half of the participants tasted all of the water drinks before the water and Tween drinks, whereas the remainder had the drinks in the opposite order. Half of the participants received the CS block before the US block, whereas the remainder received the blocks in the opposite order.

Results

Both groups found the US aversive: mean ratings of the CS+ drinks were –66.56 ($SE = 4.67$) in the experimental group and –52.40 ($SE = 5.01$) in the control group. However, the experimental group found the CS+ more aversive, $t(62) = -2.07$, $p < 0.05$, which should make conditioning effects compared with the control group stronger.

Table 2 shows the mean ratings of the drinks pre- and post-conditioning. In the experimental group, both CS+ and CS– drinks showed negative shifts in ratings across conditioning, although these shifts were more pronounced in CS– drinks. In the control group, both CS+ and CS– drinks were rated more positively after conditioning. A 2 (group: experimental or BSB control) × 2 (US type: CS+ or CS–) × 2 (phase: pre- or post-conditioning) mixed ANOVA revealed a significant effect of group, $F(1, 62) = 4.99$, $p < 0.05$, but no significant effect of US type, $F(1, 62) = 1.74$, or the phase × group interaction, $F(1, 62) =$

2.37. There were no significant effects of US type × group, phase, US type × phase, or the crucial US type × phase × group interaction, all Fs (1, 62) < 1. The conditioning effect, as shown by the three-way interaction, yielded a small effect size, $r = 0.11$.

Discussion

Two attempts to directly replicate the basic conditioning of flavour preferences both failed to demonstrate any evidence of evaluative conditioning. Despite being close replications of published work that obtained strong effects, and having comparable sample sizes to those used in the original studies, no significant conditioning effects were observed. There are several possibilities to explain these failures: (1) EC does not exist and prior successes have resulted from artefacts in the methodology; (2) the CS and US are not similar enough leading to a weak EC effect as predicted by similarity based models of EC (Davey, 1994; Havermans & Jansen, 2007); (3) the USs were not powerful enough; and (4) the CS-US pairings were not obvious to participants. With any failure to elicit an effect there could also be any number of other unmeasured factors that could make a difference. Based on self-reported ratings, the US certainly had the desired effect. Also, given the methodology closely approximated earlier studies we should be able to assume that CS-US pairings were at least as obvious in our studies as in prior ones. It is possible that our CSs were not similar enough to the US to create stimulus generalisation but we cannot ascertain this from our data. Finally, given that EC using tastes has been reported in many studies in different laboratories it seems unlikely that all of these successes are due to artefacts. However, Experiments 7 and 8 do show that the effects are not as robust as one might assume.

Conceptual similarity between the CS and US

Two theoretical models of EC have 'similarity' as a central theme (Davey, 1994; Havermans & Jansen, 2007). Early studies into EC matched CSs and USs on the basis of visual similarity. However, as we have already discussed, in doing so CSs were not counterbalanced across USs; therefore, the possibility that the conditioning effects observed resulted from the properties of the stimuli themselves rather than the conditioning procedure itself could not be ruled out (Shanks & Dickinson, 1990). This point is particularly pertinent given that subsequent research showed that conditioning-like effects that resulted from experimental artefacts could be obtained when CSs and USs were matched for perceptual similarity (Field & Davey, 1999). This presents us with a problem: a lack of similarity between CS and US could explain some of our failures to obtain EC, but to rule out artefactual explanations of EC effects we need to counterbalance CSs across USs. One solution is to look not at visual similarity, but conceptual similarity. Field (2000a) has argued that EC should occur (or at least be stronger) in CS-US pairings that have an obvious conceptual connection for the participant. This idea is consistent with Davey's (1994) suggestion that the mechanism underlying evaluative conditioning is a conceptual transfer between the CS and US. In Davey's model, the conditioning procedure creates learning by alerting participants to the common conceptual features of the CS and US. This awareness of common conceptual features leads participants to revalue the affective properties of the CS in line with the affective response evoked by the US. It is also consistent with Havermans and Jansen's (2007) stimulus generalisation model, in as much as the similarity in their model derives from shared common features (and these features need not be only visual). A strong conceptual link makes it likely that there will be more common features between the CS and US. As such Experiments 9 to 12 manipulate the conceptual similarity between the CS and US. In doing so we can create CS-US pairings that can be counterbalanced to rule out artefacts, but also test a theoretical prediction that similarity between the CS and US (in a broad sense) will enhance EC.

Experiment 9

Method

Participants. Seventeen participants completed the experiment and were paid a small sum for their participation (age $M = 25.14$, $SD = 4.33$).

Stimuli. Forty-six pictures were rated by 12 independent judges using a standard EC rating scale ranging from -100 (dislike), through 0 (neutral) to +100 (like). To try to control conceptual similarity between the CSs and USs, all of the pictures depicted scenes with several people in them. Two disliked US pictures ($Ms = -81.67$ and -86.67, $SDs = 16.83$ and 17.23), two liked US pictures ($Ms = 48.33$ and 50.00, $SDs = 31.14$ and 31.193) and four neutral CSs ($Ms = -7.08$, -3.75, –1.67, and 0.83, $SDs = 15.44$, 16.94, 18.13, and 12.40) were selected based on mean ratings, low standard deviations (implying a consistency across judges) and similarity in ratings between male and female raters. The remaining 38 pictures were used as filler stimuli.

Procedure. A similar experimental procedure was used as in Experiment 1. The pictures were evaluated, using the rating scale already described, both before and after a conditioning stage. Each participant saw 2 × Neutral-Liked CS-US pairs and 2 × Neutral-Disliked pairs. The US paired with a particular CS was counterbalanced across participants using a Latin-square method. Each pair was presented ten times during conditioning, with each stimulus appearing for 1 second, the trace interval being 200 ms, and an ITI of 4 seconds.

Results

The mean US ratings were –79.12 for the disliked USs (*SD* = 26.15) and 56.03 for the liked USs (*SD* = 29.54), which were significantly different, $t(33) = 17.10$, $p < 0.001$.

Despite the neutral CSs being rated as subjectively neutral in the pilot study, some participants rated some of these CSs either very positively (greater than +30) or very negatively (less than –35). It may be important for EC that CSs are perceived as relatively neutral for them to acquire the valence of a US, hence, data were examined both including and excluding these CSs. Exclusion of non-neutral CSs did not alter the results and so analysis for the complete dataset is reported. Table 2 shows the mean CS ratings pre- and post-conditioning for all CSs.

A two-way 2 (phase: preconditioning or post-conditioning) × 2 (US type: liked US or disliked US) related ANOVA was conducted revealing a significant main effect of phase, $F(1, 16) = 27.56$, $p < 0.001$, but no significant main effect of US type, $F(1, 16) = 1.75$, and more important no significant interaction, $F(1, 16) < 1$. The conditioning effect seen in this interaction reflects a small effect size, $r = 0.17$. In short, although ratings fell across conditioning, these changes did not depend on the US.

Experiment 10

Experiment 10 attempted to improve upon Experiment 9 by systematically varying whether CS-US pairs were conceptually linked. The conceptual link used in the experiment was whether the CSs or USs were baby relevant.

Method

Participants. Thirty-two unpaid participants completed the experiment with 16 participants assigned to both the experimental (age *M* = 29.56, *SD* = 10.65 years) and BSB control conditions (age *M* = 28.88, *SD* = 8.27).

Stimuli. Forty-three colour images were used: 41 were taken from the IAPS, and two further images were obtained by the experimenters (a picture of a bottle of milk, and a jar of baby food). Four CS pictures were used: a bottle of milk, a jar of baby food, a white bowl and a box of tissues. The negative USs were a dead person's head and vomit in a toilet bowl. The positive USs were a baby and a seal pup. These CSs and US were selected so that for each participant there could be two conceptually related CS-US pairs (one with a disliked US and one with a liked one) and two conceptually-irrelevant pairs (one with a disliked US and one with a liked). All of the CSs were chosen to be 'baby related' (i.e. a baby's food bowl, tissues, milk and baby food) and one positive US (a baby) and one negative US (the vomit) were baby related. As such, each participant received four different pairs: 2 × related pairs (baby-related CS with baby-related positive US, and baby-related CS with baby-related negative US) and 2 × unrelated pairs (baby-related CS with unrelated positive CS, and baby-related CS with unrelated US).

Procedure. Participants rated all of the pictures in random order at preconditioning and post-conditioning. In the conditioning phase, ten pairings of each of the four CS-US pairs were presented in semi-random order (random but with the restriction that no pairing could appear more than twice in succession); the CS was displayed for 2 seconds followed by a trace interval of 200 ms, then the US appeared for 2 seconds followed by an ITI of 5 seconds. The BSB control followed these timings but CSs and USs were presented separately (see Experiment 2). Finally participants engaged in an awareness measure in which each CS appeared in turn on the left of the screen (in random order) and all four US appeared on the right (in random locations). Participants had to select the US that was paired with each CS by clicking on the US and then indicate their confidence by selecting one of four screen buttons labelled 'completely sure', 'rather sure' rather unsure' and 'completely unsure'. A second awareness measure asked participants to indicate the valence of the US with which each CS was paired by selecting one of three screen buttons labelled 'liked', 'disliked' and 'neutral'.

Participants were considered contingency aware if they correctly identified three or four the pairs using the first awareness measure. In cases where participants identified less than 3 correct pairings, the data from the second awareness measure was used with participants classified as aware if they identified the correct valence of the US for three or four of the CSs.

Results

The average ratings of the USs in N-L and N-D pairs were 25.00 (*SE* = 6.74) and –78.13 (*SE* = 4.41) respectively in the experimental group, and 42.81 (*SE* = 6.74) and –66.72 (*SE* = 4.41) in the BSB control. A two-way US type (N-L, N-D) × group (experimental or BSB control) ANOVA revealed that these differences were significant, $F(1, 30) = 301.10$, $p < 0.001$: USs in N-L pairs were rated significantly higher than those in N-D pairs. There was no group × US type interaction, $F(1, 30) < 1$, indicating that this pattern of results was consistent across experimental and control participants.

Table 2 shows the mean evaluative responses for relevant and irrelevant pairings. A 2 (relevance: relevant or irrelevant) × 2 (US type: liked or disliked) × 2 (phase: pre- or post-conditioning) × 2 (group: experimental or BSB control) mixed ANOVA revealed no significant effects at all:

main effects of relevance and US type, and interactions of relevance × group, relevance × US type × group, and relevance × US type × phase, $Fs(1, 30) = 2.19, 2.69, 1.44, 1.90$, and 2.85 respectively; all other $Fs(1, 30) < 1$.

The data for 'contingency aware' and 'unaware' participants were analysed using a 2 (relevance: relevant or irrelevant) × 2 (US type: liked or disliked) × 2 (phase: pre- or post-conditioning) repeated measures ANOVA. In the six participants classified as contingency unaware there was a significant US type × phase interaction, $F(1, 5) = 7.67$, $r = 0.79$, indicating that CS ratings changed significantly across conditioning as a function of the type of US used. This effect was not significantly influenced by the relevance of the CS-US pairing. This effect was opposite to what would be expected: collapsing across CS-US relevance, ratings of CSs paired with liked USs went down (from 2.75 to -0.25), whereas ratings of those paired with disliked USs went up (from -3.00 to 0.25). In the ten participants classified as contingency aware, there was a near-significant effect of the relevance × US type × phase interaction, $F(1, 9) = 4.73$, $p < 0.06$, $r = 0.59$. This represented a medium effect size. (In fact, this interaction was significant, $F(1, 7) = 7.43$, $r = 0.72$, when data from two participants, who revealed post-experiment that they deliberately kept their ratings consistent across conditioning because they thought the experimenter wished them to do so, were excluded.) Table 2 shows that for relevant pairings the predicted changes in CS ratings were observed (those paired with liked USs went up, those paired with disliked USs decreased), for irrelevant pairings effects seemed to be weaker (ratings of CSs paired with liked USs increased but those paired with disliked USs did not change). No other effects were significant: all $Fs(1, 9) < 1$. Unfortunately these effects could not be compared with the BSB control because of awareness being measured on a post hoc basis. It is not possible to similarly divide the BSB control into aware and unaware groups (all participants should be unaware); therefore, it is not clear with which participants in the BSB control group we could compare those in the aware and unaware groups. Experiments 11 and 12, therefore, attempted to manipulate awareness of the pictures themselves to enable meaningful comparisons with a BSB control group.

Experiment 11

Experiment 10 demonstrated an unexpected interaction between contingency awareness and the effects of conditioning with contingency unaware participants demonstrating a conditioning effect in the predicted direction whereas those aware of the contingencies showed this pattern in relevant pairings, but showed a reverse response pattern to irrelevant pairings. However, contingency awareness was measured as a tangential theme, rather than being manipulated systematically, with participants classified as being only broadly aware or unaware of all contingencies. Second, because of the small sample sizes, adequate comparisons with the BSB control group could not be performed. In Experiment 11, attempts are made to make all participants aware of the contingencies, whereas in Experiment 12, contingency awareness is prevented by using backward masked subliminal presentations.

Method

Participants. Thirty-two unpaid participants completed the experiment with 16 participants assigned to both the experimental (age $M = 21.50$, $SD = 3.32$ years) and BSB control conditions (age $M = 21.88$, $SD = 2.85$).

Stimuli. Forty-three colour images from Experiment 10 were used, except the four CSs and one of the positive USs were replaced. The four CS pictures used in the present experiment were a dog-food bowl, a dog collar, a dog basket, and a can of dog food. The negative USs were a dead person's head and a picture of a dead dog. The positive USs were some puppies and a beautiful landscape. These CSs and US were selected so that for each participant there could be two conceptually related CS-US pairs (one with a disliked US and one with a liked one) and two conceptually-irrelevant pairs (one with a disliked US and one with a liked). All of the CSs were chosen to be 'dog related' (i.e. a dog bowl, dog collar, dog basket and dog food) and one positive US (puppies) and one negative US (a dead dog) were dog related. The intention was to have a stronger conceptual connection between CSs and certain USs than in Experiment 10 (the vomit US, for example, may be associated with babies in participants who are parents, but in students it might be associated more strongly with other activities). As such, each participant received four different pairs: 2 × related pairs (dog-related CS with dog-related positive CS, and dog-related CS with dog-related US) and 2 × unrelated pairs (dog-related CS with unrelated positive CS, and dog-related CS with unrelated US).

Procedure. The procedure was similar to Experiment 10 aside from the conditioning phase. During conditioning, participants were instructed to memorise the order of the pictures. This was intended to increase contingency awareness. Timing parameters were also changed to make the pairings more obvious: the CS was displayed for 1 second followed by a trace interval of 100 ms, then the US appeared for 1 second followed by an ITI of 4 seconds.

Results

Data from one participant in the experimental group were unavailable due to a technical failure. All participants in the experimental group were classified as contingency aware, whereas those in the control group (who saw no contingencies) were all classified as contingency unaware.

The average ratings of the USs in N-L and N-D pairs were 39.83 ($SE = 7.53$) and –70.33 ($SE = 7.36$) respectively in the experimental group, and 55.63 ($SE = 7.29$) and –72.81 ($SE = 7.13$) in the BSB control. A two-way US type (N-L, N-D) × group (experimental or BSB control) ANOVA revealed that these differences were significant, $F(1, 29) = 180.08$, $p < 0.001$: USs in N-L pairs were rated significantly higher than those in N-D pairs. There was no group × US type interaction, $F(1, 29) < 1$, indicating that this pattern of results was consistent in the experimental and control participants.

Table 2 shows the mean evaluative responses for relevant and irrelevant pairings. A 2 (relevance: relevant or irrelevant) × 2 (US type: liked or disliked) × 2 (phase: pre- or post-conditioning) × 2 (group (experimental or BSB control) mixed ANOVA revealed no significant main effects of US type ($F(1, 29) = 2.75$) or relevance, phase, and group ($Fs\ (1, 29) < 1$). There were also no significant interactions between relevance × group ($F(1, 29) = 2.18$), time × group, US type × group, US type × relevance, relevance × US type × phase, relevance × group × phase ($F(1, 29) = 1.68$): all $Fs(1, 29) < 1$ unless stated. There was a marginal effect of US type × phase, $F(1, 29) = 3.63$, $p = 0.067$, $r = 0.33$, and relevance × phase, $F(1, 29) = 3.57$, $p = 0.069$, $r = 0.33$. However, there was a significant group × US type × phase interaction, $F(1, 29) = 13.23$, $r = 0.56$, indicating a conditioning effect; however, this effect was not influenced by the relevance of the pairings as shown by a non-significant four-way interaction, $F(1, 29) < 1$.

This demonstrates that significant conditioning effects were found (changes in ratings to CSs paired with positive USs changed differentially in the expected direction to those paired with disliked USs) compared with a BSB control. However, the relevance of the pairings, unlike in Experiment 10, had no influence on the effects observed.

Experiment 12

The purpose of this experiment was to look at the effects of CS-US relevance in contingency-unaware participants.

Participants. Forty unpaid participants were recruited with 21 participants assigned to the experimental condition (age $M = 24.10$, $SD = 5.28$ years) and 19 participants assigned to the BSB control condition (age $M = 23.32$, $SD = 4.66$).

Stimuli. Identical stimuli were used as in Experiment 11.

Procedure. During the conditioning phase, contingency awareness was reduced by using rapidly presented backward masked US presentations, in which the US and mask appeared over a 1 second interval with the US occupying one refresh rate of the monitor (17 ms approx.) of the interval and the mask occupying the remaining 983 ms. Each US had a unique mask, consisting of a pattern of noise (random colour dots) that had either a green, blue, red or yellow filter. As such the masks had no recognisable features or similarity with the US. For each CS-US pair, the CS appeared for 1 second followed by a 100 ms trace interval and then the US/mask. The ITI was 4 seconds. The BSB control group was the same as previous experiments except that USs were still rapidly presented with their masks. Following awareness measures (as in the previous two experiments) a measure of conceptual relevance was taken to ensure that participants did regard relevant and irrelevant pairings in the way intended. In this stage each CS-US pair appeared on the screen (in random order) underneath which there was 100 point scale ranging from 0 (completely unrelated) to 100 (completely related) which participants could manipulate using the computer mouse.

Results

The average ratings of the USs in N-L and N-D pairs were 52.83 ($SE = 5.08$) and –82.95 ($SE = 3.76$) respectively in the experimental group, and 48.16 ($SE = 5.34$) and –80.58 ($SE = 3.95$) in the BSB control. A two-way US type (N-L, N-D) × group (experimental or BSB control) ANOVA revealed that these differences were significant, $F(1, 38) = 590.17$, $p < 0.001$: USs in N-L pairs were rated significantly higher than those in N-D pairs. There was no group × US type interaction, $F(1, 38) < 1$, indicating that this pattern of results was consistent in the experimental and control participants.

All participants were deemed contingency unaware. Dog-related CS-US pairings were rated as highly related for both N-L ($M = 83.67$, $SE = 3.00$) and N-D pairs ($M = 46.27$, $SE = 4.75$). Unrelated CS-US pairs were viewed as relatively unrelated for both N-L ($M = 10.06$, $SE = 3.02$) and N-D pairs ($M = 5.00$, $SE = 2.39$). A three-way US type (N-L, N-D) × relevance (dog-related or unrelated) × group (experimental or BSB control) ANOVA revealed a significant difference in relevance ratings created by the type of US, $F(1, 38) = 58.38$, $p < 0.001$, and whether the CS and US were supposed to be relevant, $F(1, 38) = 295.21$, $p < 0.001$. There was also a significant relevance × US type interaction, $F(1, 38) = 27.63$, $p < 0.001$, indicating that the relevance manipulation was more successful for N-L pairs than it was for N-D pairs. (This is probably because the negative US depicts a badly burned dog that some may not have recognised as a dog).

Table 2 shows the mean evaluative responses for relevant and irrelevant pairings. A 2 (relevance: relevant or irrelevant) × 2 (US type: liked or disliked) × 2 (phase: pre- or post-conditioning)

× 2 (group: experimental or BSB control) mixed ANOVA revealed no significant main effects of US type, relevance, phase, and group ($Fs(1, 38) < 1$). There were also no significant interactions between relevance × group ($F(1, 38) = 1.13$), phase × group, phase × relevance, US type × group, US type × phase, US type × phase × group, US type × relevance, US type × relevance × group ($F(1, 38) = 3.47$), relevance × US type × phase, relevance × group × phase: all $Fs(1, 38) < 1$ unless stated. There was a significant effect of US type × phase × relevance, $F(1, 38) = 4.27$, $r = 0.32$. This shows that the changes in CS ratings were significantly different in N-L and N-D pairs when those pairs were irrelevant, $t(39) = 1.85$, $r = 0.28$, but not when the pairs were relevant, $t(39) < 1$, $r = 0.07$. However, the US type × phase × relevance × group interaction was nonsignificant, $F(1, 38) < 1$, indicating that this effect was similarly strong in experimental and control groups. The changes observed were the opposite of what we might expect: ratings of CSs paired with irrelevant negative pictures shifted in a more positive direction, whereas those paired with irrelevant liked pictures shifted in a more negative direction. This demonstrates that although a reverse conditioning-like effect was observed for irrelevant pairings, this effect was found in the control group also and so was not due to the pairing process.

Discussion

The results of experiments 9 to 12 were inconsistent. Experiment 9 showed little evidence of EC at all. In Experiment 10, contingency unaware participants demonstrated what looked liked a conditioning effect, whereas contingency aware participants seemed to show the effect only in relevant pairings: in irrelevant pairings a reverse conditioning effect was shown. However, the post hoc allocation of participants as ‘contingency aware’ or ‘unaware’ made it impossible to compare these effects to a BSB control (in which participants can not be similarly divided). Experiments 11 and 12 sought to unravel these findings by systematically manipulating contingency awareness. Experiment 11 showed that when participants were contingency aware a conditioning effect was observed; however, Experiment 12 showed that when participants were contingency unaware there was a reverse EC effect, but only in irrelevant pairings, and this effect was not significantly greater than in a control condition. The only hint of consistency (Experiments 10 and 12) was that irrelevant pairings led to reverse effects to what would be expected in an EC experiment. Even though the process that led to these effects is non-associative this could indicate that the stimuli chosen in EC experiments can impact upon how participant’s ratings change.

General discussion

This paper has provided details of 12 EC experiments, some of which have (1) simply attempted to replicate previous successes (e.g. Experiments 1, 2, 3, 6 and 7); (2) manipulated the procedural details of the conditioning procedure to see if these act as boundary conditions, e.g. reducing the ISI (Experiments 2 and 4), reducing the number of trials (Experiment 5), increasing stimulus presentation durations (Experiment 5), reducing the number of different CS-US pairs to decrease attentional load (Experiments 3, 6, 9, 10, 11, and 12); and (3) investigated procedures that, based on theory, would be expected to enhance conditioning and associative learning, e.g. facilitating awareness of CS-US pairings (Experiments 3, 4, 11, and 12), and enhancing conceptual similarity between CS and US (Experiments 9 to 12). In the vast majority of these studies no differential EC effects were observed – despite the fact that these failures could not be attributed to small participant numbers or lack of statistical power. EC effects were not replicated across more than one modality (both visual-visual and flavour-flavour paradigms), in fully counterbalanced stimulus designs which rule out artefactual stimulus selection effects (Field & Davey, 1999), and when evaluative changes to paired stimuli are compared with nonpaired control conditions.

If boundary conditions do limit the success of EC (De Houwer et al., 2005; De Houwer et al., 2000; De Houwer et al., 2001; Rozin et al., 1998), then none of the parametric manipulations in the present studies illuminate where those boundaries may lie. Fairly basic variations in such parameters as the ISI, number of pairing trials, CS and US duration, and number of different CS-US pairs failed to facilitate EC, and while it may be argued that these variables were not manipulated in a truly systematic way, the effect sizes reported suggest that the failures were statistically legitimate. Other parametric aspects of EC that have yet to be explored might enhance conditioning (e.g. the use of simultaneous conditioning procedures, or the adoption of the simple CS+/CS- procedure usually used in human autonomic conditioning paradigms). However, simply varying the very basic parameters does not seem to tap those processes that might be central to generating successful EC.

Contingency awareness also appears to have no consistent effect: contingency awareness sometimes does not moderate EC effects, sometimes EC occurs only with contingency awareness (see Field, 2000b for a review), on other occasions contingency awareness actually prevents learning or creates contrast effects (Fulcher & Hammerl, 2001). The present experiments show similarly contrasting effects. In Experiment 3, awareness had no effect on EC whilst unaware participants showed a contrast effect regardless of CS-US relevance in Experiment 10. Aware participants showed some evidence of EC effects in relevant CS-US pairings in Experiment 10. EC effects were apparent in both relevant and irrelevant pairings in Experiment 11 in which all participants were able to consciously perceive the stimulus presentations. However, in Experiment 12 in which stimulus presentations were subliminal, EC effects were observed only for irrelevant pairings! In most of these cases effects were not significantly different to those in the BSB control. At best, therefore, we could conclude that CS-US relevance interacts with contingency awareness in some way, but that future research is needed to unpack this relationship (and that this relationship might be due to nonassociative factors).

Conceptual similarity between the CS and US was also investigated. This variable seems like a good candidate as a moderating factor (or boundary condition) for several reasons. First, there is evidence from Pavlovian conditioning that effects such as conditioning without contingency awareness and resistance to extinction that are argued to be unique to EC can be found when the learning episode has ecological relevance (Field, 2000a). This leads to the idea that EC will be strongest when the CS and US make conceptual sense (Field, 2000a, 2001). Similarly, De Houwer et al. suggested that 'belongingness' between the CS and US might be an important moderator of EC. This is partially supported by the fact that the evidence for cross-modal conditioning appears to be somewhat weaker than for intra-modal conditioning, which of course is also a prediction from Havermans and Jansen's (2007) model. There is also some empirical evidence that expectancies about whether the CS and US go together affects EC. Lascelles, Field & Davey (2003) suggest that if participants expect two things to go together then an association is formed more easily between them. They found that evaluations of foodstuffs were changed using an EC paradigm only when paired with certain kinds of body shapes. Finally, conceptual similarity as a moderator of EC effects fits with two theoretical models (Davey, 1994; Havermans & Jansen, 2007), both of which suggest that similarity between the CS and US should facilitate conditioning. In Havermans and Jansen's model especially, conceptual similarity relates to the number of common features (not necessarily visual features) that a participant can identify in the CS and US, which in turn affects that formation of a configural representation of both stimuli. In Experiments 10 to 12, by using baby-related/unrelated and dog-related/unrelated pairings the number of common conceptual features that the CS and US shared were manipulated. Based on their model, EC should have been strongest for the 'related' pairings. Unfortunately, the results interacted with contingency awareness and were less than clear cut. Experiment 10 showed stronger effects in conceptually relevant CS-US pairings in participants aware of the contingencies. In unaware participants, CS-US relevance did not have a significant effect; however, the means suggested a stronger effect in related pairings and the sample size ($n = 6$) was likely to be too small to detect the effect of relevance. In any case, these results need to be viewed with caution. When contingency awareness was systematically manipulated in Experiment 11, CS-US relevance did not moderate EC effects. However, the dimension of CS-US relevance changed from 'babies' to 'dogs' in this experiment, which could explain the differences to Experiment 10 (perhaps 'babies' is a stronger conceptual connection than 'dogs' or the conceptual relevance was clearer in the baby stimuli compared with the dog stimuli). In Experiment 12, in which contingency awareness was hampered by subliminal presentation of stimuli, there appeared to be an effect for unrelated CS-US pairings, but this was a contrast effect (the changes in CS evaluations were opposite to the US with which they were paired) and was not significantly different from the control group. All in all, the best we can (very tentatively) conclude is that CS-US relevance might facilitate EC in contingency aware participants. This conclusion would support similarity-based models of EC (Davey, 1994; Havermans & Jansen, 2007) but clearly more work needs to be done, and the similarity between the CS and US needs to be more specifically defined.

A final factor to consider is attentional load, which is an important feature of models of Pavlovian conditioning generally. The Rescorla-Wagner model (1972) formalises the idea that associations are formed between cues and surprising outcomes. This model famously incorporates a term representing a cue's individual associability, which represents an individual learning rate that the model acknowledges stems from differential attention. The attention devoted to a given cue is a function of its importance in predicting an outcome: that is, animals will attend to relevant stimuli at the expense of not attending to irrelevant ones (Mackintosh, 1975). Both models formalise learning in terms of a change in the association weights (associative strength) of a CS. This associative strength is also an important component of Pearce's (1987) stimulus generalisation model and so attentional load (as something that impacts upon associative strength) is important in Havermans and Jansen's (2007) model of EC: as attentional load decreases, EC should increase. Four experiments (Experiments 1, 2, 4 and 5) used 9 CS-US pairs, whereas another five (Experiments 9, 10, 11, 12, and 6) used only 4. Comparing the effect sizes in these studies using Hunter and Vevea's (1998) random effects meta-analysis and the number of CS-US pairs as a moderator variable (Field & Gillett, 2007), shows that the moderating effect of the number of pairs on the effect size is significant and represents a medium effect, $\beta = 0.3$, $\chi^2 = 3.68$, p (one-tailed) < 0.05. This suggests that a large attentional load reduces EC effects. This is consistent with Field and Moore (2005) who showed that distracter tasks can prevent EC, and with Davey's (1994) suggestion that EC might be difficult to obtain when relatively large numbers of different CS-US pairings are used, which inflict a heavy processing load on participants. It is also consistent with Havermans and Jansen's (2007) model of EC, which predicts precisely this effect because attention is linked to associative strength (e.g. Mackintosh, 1975; Rescorla & Wagner, 1972) and, therefore, when lots of CS-US pairs are used some strong CS-US associations might be made to CSs that attract attention at the expense of attention to other CSs: when averaged out, these strong and weak associations cancel out.

Whilst it is informative to look at large bodies of work that have manipulated various aspects of the EC paradigm, easy answers are hard to find. While there is tentative evidence for effects of attentional load and perhaps an interaction between conceptual similarity and contingency awareness, more work is required. In particular it may be worth applying some of the principles from established models of associative learning to further understanding. For example, we know from Rescorla-Wagner that the associability of a CS is important, as is the effectiveness of the US and the amount of associative value that remains to be conditioned. None of these variables have really been explored: we assume very negative and positive images will be equally as effective as USs, and we assume that neutral CSs are equally associable and have equal associative value remaining. This may not be the case. Systematic studies that investigate these sorts of factors are important, even if they yield non-significant effects, if we are truly to understand the boundary conditions of EC.

Acknowledgement

This research was funded by Unilever Plc. grant RBW3 0600 to Andy Field, and by grants from the Biolotechnical and Biological Sciences Research Council (award number 85/S10580) and the Economic and Social Research Council (R000235939) to Graham Davey. Thanks to Tara Said (Experiment 6), Anna Wittington (Experiment 7), and Matt Knight (Experiment 8), who helped collect data.

References

Baeyens, F., Crombez, G., Hendrickx, H. & Eelen, P. (1995). Parameters of Human Evaluative Flavor-Flavor Conditioning. *Learning and Motivation, 26*(2), 141-160.

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.

Baeyens, F. & De Houwer, J. (1995). Evaluative Conditioning Is a Qualitatively Distinct Form of Classical-Conditioning - Reply. *Behaviour Research and Therapy, 33*(7), 825-831.

Baeyens, F., Eelen, P. & Crombez, G. (1995). Pavlovian Associations Are Forever - on Classical-Conditioning and Extinction. *Journal of Psychophysiology, 9*(2), 127-141.

Baeyens, F., Eelen, P., Crombez, G. & Vandenbergh, O. (1992). Human Evaluative Conditioning - Acquisition Trials, Presentation Schedule, Evaluative Style and Contingency Awareness. *Behaviour Research and Therapy, 30*(2), 133-142.

Baeyens, F., Eelen, P. & van den Bergh, O. (1990). Contingency Awareness in Evaluative Conditioning - a Case for Unaware Affective-Evaluative Learning. *Cognition & Emotion, 4*(1), 3-18.

Baeyens, F., Eelen, P., van den Bergh, O. & Crombez, G. (1990). Flavor-Flavor and Color-Flavor Conditioning in Humans. *Learning and Motivation, 21*(4), 434-455.

Baeyens, F., Hermans, D. & Eelen, P. (1993). The Role of CS-US Contingency in Human Evaluative Conditioning. *Behaviour Research and Therapy, 31*(8), 731-737.

Davey, G. C. L. (1994). Defining the Important Theoretical Questions to Ask About Evaluative Conditioning - a Reply. *Behaviour Research and Therapy, 32*(3), 307-310.

De Houwer, J., Baeyens, F. & Field, A. P. (2005). Associative learning of likes and dislikes: Some current controversies and possible ways forward. *Cognition and Emotion, 19* (2), 161-174.

De Houwer, J., Baeyens, F., Vansteenwegen, D. & Eelen, P. (2000). Evaluative conditioning in the picture-picture paradigm with random assignment of conditioned stimuli to unconditioned stimuli. *Journal of Experimental Psychology-Animal Behavior Processes, 26*(2), 237-242.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*(6), 853-869.

Diaz, E., Ruiz, G. & Baeyens, F. (2005). Resistance to extinction of human evaluative conditioning using a between-participants design. *Cognition and Emotion, 19*(2), 244-268.

Field, A. P. (1997). *Re-evaluating evaluative conditioning.* University of Sussex, Brighton, UK.

Field, A. P. (2000a). Evaluative conditioning is Pavlovian conditioning: Issues of definition, measurement, and the theoretical importance of con-

tingency awareness. *Consciousness and Cognition, 9*(1), 41-49.
Field, A.P. (2000b). I like it, but I'm not sure why: Can evaluative conditioning occur without conscious awareness? *Consciousness and Cognition, 9*(1), 13-36.
Field, A.P. (2001). When all is still concealed: Are we closer to understanding the mechanisms underlying evaluative conditioning? *Consciousness and Cognition, 10*(4), 559-566.
Field, A.P. (2005). Learning to Like (and dislike): Associative Learning of Preferences. In A.J. Wills (Ed.), *New Directions in Human Associative Learning* (pp. 221-252). Mahwah, New Jersey: LEA.
Field, A.P. & Davey, G.C.L. (1997). Conceptual conditioning: Evidence for an artifactual account of evaluative learning. *Learning and Motivation, 28*(3), 446-464.
Field, A.P. & Davey, G.C.L. (1998). Evaluative conditioning: Arti-fact or -fiction? A reply to Baeyens, de Houwer, Vansteenwegen, and Eelen (1998). *Learning and Motivation, 29*(4), 475-491.
Field, A.P. & Davey, G.C.L. (1999). Reevaluating evaluative conditioning: A nonassociative explanation of conditioning effects in the visual evaluative conditioning paradigm. *Journal of Experimental Psychology-Animal Behavior Processes, 25*(2), 211-224.
Field, A.P. & Gillett, R. (2007). How to do a meta-analysis. *Manuscript Under Review*.
Field, A.P. & Moore, A.C. (2005). Dissociating the effects of Attention and contingency awareness on evaluative conditioning effects in the visual paradigm. *Cognition and Emotion, 19*(2), 217-243.
Fulcher, E.P., & Hammerl, M. (2001). When all is revealed: A dissociation between evaluative learning and contingency awareness. *Consciousness and Cognition, 10*(4), 524-549.
Havermans, R.C. & Jansen, A. (2007). Evaluative conditioning: a review and a model. *Netherlands Journal of Psychology, 63*(2), 38-49.
Hedges, L.V. & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486-504.
Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1997). *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. Florida: NIMH Center for Emotion and Attention (CSEA).
Lascelles, K.R.R., Field, A.P. & Davey, G.C.L. (2003). Using foods as CSs and body shapes as UCSs: A putative role for associative learning in the development of eating disorders. *Behavior Therapy, 34*(2), 213-235.
Mackintosh, N.J. (1975). Theory of Attention - Variations in Associability of Stimuli with Reinforcement. *Psychological Review, 82*(4), 276-298.
Martin, I. & Levey, A. (1994). The Evaluative Response - Primitive but Necessary. *Behaviour Research and Therapy, 32*(3), 301-305.
Pearce, J.M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, 61-73.
Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and on reinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.
Rozin, P., Wrzesniewski, A. & Byrnes, D. (1998). The elusiveness of evaluative conditioning. *Learning and Motivation*, 29(4), 397-415.
Shanks, D.R. & Dickinson, A. (1990). Contingency Awareness in Evaluative Conditioning - a Comment. *Cognition & Emotion*, 4(1), 19-30.