Evaluative Conditioning and the Awareness Issue: Assessing Contingency Awareness With the Four-Picture Recognition Test

Eva Walther and Benjamin Nagengast
University of Heidelberg

An experiment is described that tested the moderating influence of contingency awareness on evaluative conditioning. After participants were conditioned within the picture–picture paradigm, contingency awareness was assessed by means of a recognition test (i.e., the 4-picture recognition test). Results indicate an inverse relationship between the conditioned affective reaction and contingency awareness: Only participants classified as unaware in the recognition test showed significant effects of evaluative learning. A closer inspection indicates that aware individuals stored not only the valence but also the nominal stimulus in mind.

Keywords: evaluative conditioning, awareness, recognition

Assessing Contingency Awareness in Evaluative Conditioning

Although there is substantial agreement among conditioning researchers that the question of awareness plays a crucial role in explaining evaluative learning (Baeyens, Hermans, & Eelen, 1993; Dawson & Reardon, 1973; Field, 2000, 2001; Field & Moore, 2005; Hammerl, 2000; Lovibond & Shanks, 2002), there is less agreement on how awareness should be assessed. As frequently noted, however, it is clear that the way contingency awareness is measured strongly determines whether an individual is categorized as aware or unaware.

Two different methods are chiefly used in the evaluative conditioning literature to test awareness: (a) awareness questionnaires that assess verbalizable knowledge and (b) recall or recognition measures that ask participants to identify the unconditioned stimulus (US) following contingently upon the conditioned stimulus (CS). As extensively pointed out elsewhere (De Houwer et al., 2001; Field, 2000; Lovibond & Shanks, 2002; Walther, Nagengast, & Trasselli, 2005), both measurements have their disadvantages.

In order to show that two dependent variables (in this case, tests of conscious knowledge and task performance) relate to dissociable underlying systems, we must be able to show that our test of awareness is sensitive to all of the relevant knowledge. (p. 380)

Following Shanks and St. John’s argument, it can be hypothesized that most studies have underestimated the level of awareness, simply by using insensitive verbal awareness questionnaires. As all postconditioning measures of awareness might underestimate awareness in any case because of forgetting, it is especially important that awareness tests are able to capture these declining
traces. For instance, if the CS is given and the US is supposed to be selected out of all other stimuli that were also presented in the study (e.g., Baeyens, Eelen, & Van den Bergh, 1990; Field, 2000), it may often be hard to make the correct decision because other stimuli might also be accessible, but for different reasons (e.g., because of high frequency of presentation and/or similarity). Another issue is that identifying the correct US is often not distinguished from selecting a stimulus of the same valence (e.g., Baeyens, Eelen, Crombez, & Van den Bergh, 1992; Baeyens et al., 1993; Field, 2000; Fulcher & Hammerl, 2001). If participants were not able to distinguish between the nominal stimulus and a stimulus of the same valence, this would indicate that the mere valence of a US is stored in memory during evaluative conditioning. If, however, individuals were able to discriminate the correct US from other USs, this would suggest that additional features other than valence are stored in memory.

Even more questionable are measurements in which the original CS–US pairs are presented intermixed with pairs containing the genuine CS and a decoy US (e.g., Field, 2000). From an evaluative conditioning point of view, the presentation of these pairs may strongly affect the acquired contingencies between CS and US and therefore contingency awareness.

In their review of the conditioning literature, Lovibond and Shanks (2002) criticized the current awareness measurements and decried the fact that little effort has been made to design valid awareness measurements. Dawson and Reardon convincingly argued in 1973 that some awareness tests are clearly superior to others. Comparing different awareness methods with respect to their construct validity and their sensitivity, these authors concluded that a brief recognition-based test should be preferred over other awareness tests (i.e., long recall, short recall, or long recognition test) to assess postconditional contingency awareness.

In this article, we address the awareness issue by constructing a more sensitive short recognition test to assess contingency awareness in an evaluative conditioning paradigm. More specifically, we relied strongly on Shanks and St. John’s (1994) sensitivity criterion and on Dawson and Reardon’s (1973) advice in constructing our recognition measurement.

The Four-Picture Recognition Test

In the standard form of the four-picture recognition test, participants are confronted with the CS and are asked to select the contingent US out of four other pictures. More specifically, the CS is shown on the screen and participants are asked “to select the stimulus that followed this stimulus” out of four standardized alternatives that were also presented on the screen. Similar to other recognition tests, the four-picture recognition test is supposed to assess postconditioning contingency awareness to address the critical relationship between awareness and evaluative conditioning. In the present test, however, the number of alternatives to the correct US provided by the test is reduced to a minimum. According to Shanks and St. John’s (1994) sensitivity criterion, the learning and task context should match. The fewer the stimuli that are presented in each test trial, the more similar the test becomes to the learning situation, in which only the US follows the CS. Thus, it could be assumed that the four-picture test provides a context that matches the sensitivity criterion better than other tests, in which a higher number of experimental stimuli are presented. Moreover, the four-picture test is sensitive to relevant knowledge, because it allows us to distinguish between participants able to identify the correct US and those merely selecting pictures with the same valence (e.g., Baeyens et al., 1990).

Therefore, the hypothesis that mere valence is stored could be tested against the alternative idea that individuals within an evaluative learning study acquired an association between two nominal stimuli. In the four-picture recognition test, each of the alternative stimuli controls one dimension of the correct US. To examine whether participants would be able to differentiate between the nominal source of valence (i.e., the correct US) and another stimulus of the same valence, we presented participants with the correct US and another test US of the same valence. For instance, if participants were presented with a CS and the US− of Series A, the US+ of Series B served as a control stimulus for the valence dimension in the recognition test. Moreover, we introduced a test stimulus of the opposite valence from the conditioning phase to investigate participants’ ability to differentiate between different valences. A neutral test stimulus, paired with another neutral control stimulus during conditioning, completed the test. In summary, participants were shown four stimuli, which were all presented at the same rate and preceded by a CS in the conditioning phase. Thus, the stimuli differed only with respect to valence or with respect to the particular CS with which they were paired. The position of the four pictures and the order of the test trials were selected at random for each participant. The rationale behind this stimulus selection in the test is as follows: If participants are aware of the contingency between the CS and the nominal US, they should be able to identify the correct US if the CS is given. If participants stored only the valence of the US, they should select the correct US at the same rate as the test stimulus that shares the valence with the correct US. If, however, participants encoded only that the US elicits an affective reaction at all, the valenced stimuli should be selected with the same frequency. And last but not least, if participants rely on simple guessing, they should choose all stimuli at the same rate (25%).

Method

We first tried to demonstrate evaluative conditioning with appetitive (US+) as well as aversive (US−) stimuli. After conditioning, participants were provided with the four-picture recognition test.

Participants and Design

Thirty-six female participants were recruited at a shopping mall near the University of Heidelberg. Participants were asked to take part in a study “on watching and evaluating pictures of human faces” and participated in small groups of up to 4 persons. They were compensated with a bar of chocolate and a small bottle of champagne. Participants were randomly assigned to a 3 (US: appetitive vs. control vs. aversive) × 2 (series: A vs. B) × 2 (order: recognition test or evaluation test first) design with within-subject variation on the first two factors and between-subjects variation on the last factor. Because there was no influence of the series (F = 0.53, ns), data were collapsed over this factor. Data of 1 participant were excluded from further analysis because this individual did not follow instructions. The data of 2 participants were excluded from further analysis because these participants did not select the required number of neutral stimuli in the baseline. Accordingly, 33 participants were subject to the final analysis.
Materials and Procedure

Stimuli and material were essentially the same as in previous studies (Walther, 2002; Walther & Grigoriadis, 2004). Participants were greeted by a female or male experimenter and seated in front of a computer screen. The experiment consisted of three sequential phases that were guided entirely by a computer program: the baseline phase, the conditioning phase, and the test phase. Participants were asked to put on headphones and start the program of a digitalized auditory instruction and to close their eyes in order to concentrate better on the words. The auditory instructions informed participants that it was their task to watch and sometimes to judge different types of stimuli randomly selected by the computer program (see the Appendix).

After participants had finished listening to the tape, the computer program continued. The next set of instructions appearing on the screen informed participants that it was their task to watch and sometimes judge different types of stimuli randomly selected by the computer program. Participants were asked to express their first, immediate, and spontaneous reaction toward the presented photographs on a 20-cm long graphic rating scale (labeled disliked on the left and liked on the right) by positioning the cursor on any point of the scale and then pressing the left mouse key. To avoid response sets, the graphic scale consisted of no additional numbers or other numerical labels. The PC program computed the negative judgments on the left side from negative (−1) to extreme (−100) and the positive judgments on the right side from positive (+1) to extreme (+100). The neutral (0) midpoint of the scale served as a starting position for each judgment.

Baseline

In all conditions, 131 black-and-white pictures of White male faces (approximately 7 cm wide × 8 cm long) selected from German magazines and from the Internet were displayed for 3 s, each in a full frontal view, in the center of the screen. Care was taken to select a large variety of these pictures with respect to age, expression, and attractiveness. Pretesting and postexperimental interviews revealed that none of the faces were familiar to the participants. After 3 s, the picture disappeared, the rating scale appeared, and participants had the opportunity to evaluate the target. The computer program categorized the most liked stimuli (> +80) as US’s, the pictures around zero (> −20 and < +20) as CSs, and the most disliked faces from the baseline (−80) as US’s and randomly selected the stimuli for the conditioning phases from these three categories. For each participant, 12 stimuli (8 neutral, 2 disliked, and 2 liked faces) were selected from the baseline. These 12 stimuli were necessary to form two stimulus series, A and B, each consisting of a neutral–liked, a neutral–disliked, and a neutral–neutral pair. The baseline was followed by a 5-min unrelated paper-and-pencil distracter task in which participants were requested to rate their current mood state on several adjective scales.

Conditioning Phase

The baseline phase was followed by the conditioning phase. Participants were instructed that “all you have to do now is to watch the presented faces.” In this phase, participants were presented with two series of pictures, A and B, intermixed with each other. Each series consisted of three pairs of stimuli that were each presented five times, resulting in a fixed randomized order of 30 stimulus pairs. A trace conditioning procedure, similar to previous studies (Walther, 2002; Walther & Grigoriadis, 2004), was used: Each picture was displayed for 1 s with a trace interval of 1 s (i.e., the interval between the end of the first stimulus and the beginning of the second stimulus of a pair) and an intertrial interval of 2 s (intertrial interval = offset of last stimulus of previous pair to onset of first stimulus of next pair).

We expected that the contingent presentation of the CS with a valenced (i.e., appetitive or aversive) stimulus would lead to an evaluative shift of the CS in a positive or negative direction, respectively. The conditioning phase was followed by the test phase in one condition and by the recognition test in the other.

Test Phase

Using the same procedure as in the baseline, participants were required to judge for a second time all the critical stimuli (CSs, USs) that had been presented in the conditioning phase. Difference scores (test ratings − baseline ratings) were calculated to measure evaluative changes in the judgment of the CSs. In the recognition test, participants were presented with the CS in its original size on the left side of the screen and with four smaller pictures on the right. The instruction below the pictures asked participants to select the small picture that most frequently followed the large picture. One of these pictures was the correct US, and a second picture was of the same valence as the correct US. Because each series, A and B, consisted of only one single appetitive or aversive US, the picture with the same valence as the correct US was always drawn from the opposite series. The third alternative was a US of the opposite valence as the correct US randomly selected from either the same or the opposite series. A fourth neutral stimulus, which was presented in the conditioning phase, randomly selected from either the same or the opposite series, completed the test. Thus, participants were shown four stimuli, which were all presented at the same rate and preceded by a CS in the conditioning phase. The position of the four pictures and the order of the test trials were selected at random for each participant.

Results and Discussion

Conditioning Effects

Evaluative conditioning effects were analyzed by comparing difference scores for CS− and CS+ with the neutral control stimulus. Results indicated that the repeated pairing of a neutral stimulus with a positively evaluated picture evoked a positive attitude toward the formerly neutral face (M = 3.36), whereas the pairing with the disliked US led to a negatively evaluated picture (M = −9.78), with the neutral stimulus paired with another neutral picture lying between these two scores (M = 0.68). A 3 (US: appetitive vs. control vs. aversive) × 2 (order: recognition test or evaluation test first) analysis of variance with repeated measurement on the first factor revealed a significant main effect of US valence, F(2, 32) = 3.59, p < .03, indicating that the repeated pairing of a neutral CS with a positively evaluated US evoked a positive attitude toward the CS, whereas the repeated pairing of a neutral CS with a negatively evaluated US led to a negative attitude toward the CS. No other main or interaction effect reached statistical significance (Fs < 0.93). This latter finding clearly indicated that the conditioning effect was not driven by enhanced memory triggered by a previously applied recognition test.

Awareness Effects in the Four-Picture Recognition Test

Apart from this successful replication of evaluative conditioning, we investigated the findings in the recognition test. Table 1 shows the stimuli chosen in the four-picture test across Series A and B. As indicated by the frequency distribution, an average of 41% of the participants were able to identify the correct US out of the four test stimuli. This difference between the hit and the false alarm rates was significant in both test stimuli distributions as well as in the pooled test data (see Table 1). However, if the correct stimulus was not identified, participants preferred all other test
stimuli nearly to the same degree, χ²(2, N = 33) = 2.519, p < .28, ns, indicating that participants clearly did not mentally store merely the valence of the stimulus. Had this been the case, the stimulus with the same valence should have been selected at the same rate as the nominal stimulus.

In a subsequent analysis, we computed conditioning effects for those participants who recognized two of four correct stimuli in the recognition tests (defined as “aware”) and for those who failed to recognize at least two correct US (defined as “unaware”). A 3 (US: appetitive vs. control vs. aversive) × 2 (awareness: aware vs. unaware) analysis of variance with repeated measurement on the first factor revealed a significant main effect of US valence, F(2, 62) = 2.97, p < .05, and a marginally significant interaction effect, F(2, 62) = 2.63, p < .08, indicating that awareness indeed influenced evaluative learning. Separate analysis for aware and unaware participants revealed a significant effect only for unaware participants, F(2, 36) = 7.08, p < .003, but not for participants who performed well in the recognition test, F(2, 26) = 0.13, ns. That means that contingency awareness as assessed with the four-picture test was negatively related to evaluative conditioning effects. Consistent with previous research (Fulcher & Hammerl, 2001; Walther, 2002; Walther & Trasselli, 2003), conscious awareness of the contingencies between the CS and the US did not help but rather hindered evaluative learning. A similar pattern of results emerged when the awareness criterion was rendered more difficult. Defining only those participants as aware who were able to identify three out of four correct USs yielded a significant effect only for unaware participants, F(2, 44) = 4.40, p < .01, but not for participants who performed well in the recognition test, F(2, 18) = 1.11, ns. Again, the effects shown by this analysis strongly indicated that only participants who were not able to identify the correct US in our test exhibited significant changes in the linking of the CSs.

The four-picture recognition test—apart from classifying participants as aware or unaware of the contingencies and aside from its strong relationship to the conditioning outcome—also provides insight into the cognitive mechanisms that underlie false responses. The analysis of the frequency distribution (see Table 1) revealed that participants were not able to differentiate valenced from nonvalenced stimuli if they could not identify the correct US. Because all control stimuli were presented with the same frequency, all control pictures were selected at nearly the same rate. With respect to the false alarms, only a minority selected the stimulus with the same valence on a descriptive level. Thus, it is quite unlikely that participants stored only the valence of the stimuli. If they had done so, we would have obtained a higher confusion rate between the actual US and the stimulus with the same valence.

### Table 1

<table>
<thead>
<tr>
<th>Test stimulus</th>
<th>CS⁺</th>
<th>CS⁻</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct US</td>
<td>26</td>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>US same valence</td>
<td>8</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>US different valence</td>
<td>15</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>Control</td>
<td>17</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>Σ</td>
<td>66</td>
<td>66</td>
<td>132</td>
</tr>
<tr>
<td>χ²(3, N = 33)</td>
<td>10,000</td>
<td>13,758</td>
<td>21,515</td>
</tr>
<tr>
<td>p</td>
<td>&lt;.019</td>
<td>&lt;.003</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

**Note.** The sum of frequencies in Series A and B for CS⁺ and CS⁻ is shown. Correct US = stimulus that was actually presented along with the CS; US same valence = control stimulus of the same valence presented with the same frequency as the correct US; US different valence = control stimulus of the opposite valence presented with the same frequency as the correct US; Control = control stimulus of neutral valence presented with the same frequency as the correct US.

General Discussion

Even though research on evaluative conditioning is a growing area in experimental psychology (De Houwer, Baeyens, & Field, 2005; De Houwer et al., 2001), many problems remain unresolved. Apart from the question concerning the content of learning, some researchers have suggested that the question of awareness is the most important issue. The experiments reported here extend prior research by demonstrating evaluative conditioning in unaware participants. Although Fulcher and Hammerl (2001) previously provided evidence for a dissociation between evaluative learning and contingency awareness, some researchers expressed doubts as to whether their awareness measurement was sensitive enough (cf. De Houwer, 2001). Shanks and St. John (1994) argued that the retrieval context of the performance and the awareness test should be as similar as possible. However, in Fulcher and Hammerl’s studies, participants were asked to evaluate the CS on a rating scale in the performance test, whereas during the awareness measurement they were required to recall the US paired with the CS before being asked to indicate the valence of the US. This procedure was criticized not only because of the apparent dissimilarity of the two test procedures but also because the awareness measurement could be assumed to be particularly difficult given the haptic stimuli presented in these studies (cf. De Houwer, 2001). This problem, along with the methodological issue that the authors did not control for the order of awareness test and performance test, makes it hard to interpret their otherwise highly interesting findings.

The present study contributes to the current awareness debate by demonstrating evaluative learning in unaware individuals when the order of performance measurement and awareness test is controlled and when participants are classified on the basis of a presumably more sensitive awareness test. The first examination of the four-picture recognition test revealed encouraging results. With respect to the criteria described by Lovibond and Shanks (2002), floor and ceiling effects were not observed in the test (between 30% and 50% were classified as aware). With respect to sensitivity, the four-picture recognition test measurement is therefore presumably more sensitive than verbal awareness questionnaires given postconditioning, in which usually hardly any participants can be categorized as aware (Walther et al., 2005).

Moreover, as for the question of validity, the results of the test were assessed immediately after conditioning involving the same stimuli, presumably did not influence learning, and showed a straightforward relationship to the performance test. The results obtained in our study strongly support the notion that only unaware participants exhibit effects of evaluative conditioning, whereas aware participants do not. This means that awareness of the contingency as measured with our test was not a necessary precondition for evaluative learning to occur, but rather it blocked or inhibited in some way evaluative conditioning. As Hayes and...
Broadbent (1988) suggested, two independent systems can be assumed to be involved in learning: (a) an unconscious system that accumulates predictive events in the environment and (b) a conscious hypothesis-testing system. According to this model, and in line with various other dual-process models (Chaiken & Trope, 1999), the conscious system is strongly dependent on cognitive resources, whereas the unconscious system would be independent of any such limited-capacity memory system. The hypothesis that evaluating conditioning fits well into the model of unconscious learning is supported by studies showing stronger conditioning effects in distracted participants (Fulcher & Hammerl, 2001; Walther, 2002; Walther & Trasselli, 2003).

Moreover, the present study is informative as to the content of learning. Given the nearly equal frequency distribution of choices obtained in the four-picture recognition test when individuals were not able to identify the correct US, it seems unlikely that participants stored the mere valence of the stimuli. Had that been the case, we would have found a higher rate of confusion with the same-valence control stimuli. However, we found no differences between the selection of the same-valence stimuli and the other control pictures. Taking both the recognition as well as the conditioning data into account, the present findings suggest that two different processes may be involved in affective learning. If participants are aware of the contingency between the CS and the nominal US, this specific type of awareness presumably elicits control mechanisms that inhibit the expression of the affective reactions. If, however, individuals are unable to identify the nominal US, simple associative processes dominate the expression of (dis)liking. However, it should be noted that this conclusion rests on only one study and might be premature. Given the often-noted importance of the topic (De Houwer et al., 2001), we hope that further research will elaborate on awareness measurements that help to resolve the puzzle of contingency awareness in evaluative conditioning.

References


Appendix

Digitized Auditory Instructions Heard by Participants

Welcome to our study! This study is concerned with the impression we form of other people. Psychological research indicates that any individual we encounter elicits an instinctive reaction of like or dislike. That means we know immediately whether we like a person or not. These emotional reactions occur involuntarily and automatically, that is, we are not able to control these feelings of like or dislike. The present study is directly concerned with these spontaneous feelings. We all have such emotions and they are an important piece of information. In our daily life, however, politeness often demands that we control such feelings. As a result, we often do not listen to our feelings and emotions. But that is exactly the task in this study. You will now encounter a number of individuals. Sometimes you are asked to say which feelings and reactions you have towards a person, that is, you are asked to rate the likeability of a person. This response requires only that you trust in your spontaneous feelings and emotional reactions. That is, it is only important that you give way to your feelings. Psychology also indicates that our emotional reactions are very differentiated. Like a piano that has very high and very low notes, but also many tones in between, our feelings range from strong dislike to enthusiastic like, with finely graded levels in between. In our study, you are asked to express this spectrum of emotional reactions on a rating scale graph. Please use the entire scale for expressing your like or dislike, respectively. Note that the best response is always your spontaneous emotional reaction.