RUNNING HEAD: OLSON & FAZIO (2001) REGISTERED REPLICATION

Incidental Attitude Formation via the Surveillance Task: A Pre-Registered Replication of Olson and Fazio (2001)

Tal Moran1\*, Sean Hughes1\*, Ian Hussey1, Miguel A. Vadillo2,Michael A. Olson3,

Frederik Aust5, Karoline Bading5, Robert Balas4, Taylor Benedict5, [Olivier Corneille](https://www.sciencedirect.com/science/article/pii/S2214635018301795%22%20%5Cl%20%22%21)6, Samantha B. Douglas7, Melissa J. Ferguson8, Katherine A. Fritzlen3, Anne Gast5, Bertram Gawronski9, Tobias Heycke10, Fabia Högden5, Mandy Hütter11, Benedek Kurdi12, Adrien Mierop6, Jasmin Richter5, Justyna Sarzyńska4, Colin Tucker Smith7, Christoph Stahl5, Philine Thomasius5, [Christian Unkelbach](https://www.sciencedirect.com/science/article/pii/S0022103117308685#!)5

& Jan De Houwer1

### 1Ghent University, 2Universidad Autónoma de Madrid, 3University of Tennessee, 4Institute of Psychology of the Polish Academy of Sciences, 5University of Cologne, 6Université catholique de Louvain, 7University of Florida, 8Cornell University, Department of Psychology, 9University of Texas at Austin, 10GESIS, Mannheim, Germany, 11Eberhard Karls Universität Tübingen, Department of Psychology, 12Harvard University, Department of Psychology.

**\*** TM and SH are joint first author on this paper

Abstract

Evaluative conditioning (EC) is one of the most widely-studied and highly applicable pathways for establishing and changing attitudes. Although many paradigms have been used to study EC, the surveillance task (Olson & Fazio, 2001), is one of the most cited, and purportedly generates attitudes that are claimed to occur without awareness. The ability for EC effects to occur without awareness continues to fuel conceptual, theoretical, and applied developments. Yet few published studies have used this task, and those that do are characterized by small samples and very small effect sizes. Our paper represents a high-powered (N = XXXX), pre-registered replication attempt designed to provide a strong test of EC effects produced by this task. We [did/did not] detect evidence for EC effects. Our findings [support/place strong constraints] on the surveillance task’s utility when making theoretical claims related to attitudes and EC, and its use in applied settings.

*Keywords: Pre-Registered Replication; Evaluative Conditioning; Contingency Awareness; Recollective Memory; Attitude Formation*

Incidental Attitude Formation via the Surveillance Task: A Pre-Registered Replication of Olson and Fazio (2001)

Evaluative conditioning (EC) is a widely-studied and highly applicable pathway for establishing and changing attitudes (e.g., De Houwer, Thomas, & Baeyens, 2001). In a typical EC task, a neutral (conditioned) stimulus (CS) is repeatedly paired with a positive or negative (unconditioned) stimulus (US), and as a result, the former typically acquires a similar valence as the latter.

Evaluative conditioning plays a central role in theory and application throughout psychological science. For instance, in its original version, the Associative–Propositional Evaluation (APE) Model (Gawronski & Bodenhausen, 2006), an influential theory of attitudes in social psychology, distinguished between explicit attitudes and implicit attitudes, and treated EC as a key pathway for changing the latter. The Elaboration-Likelihood Model, in the domain of persuasion (ELM; Petty & Cacioppo, 1986), distinguishes between the central and peripheral routes to persuasion, and views EC as highly relevant to the latter route. Elsewhere, EC is said to play an important role in implicit bias (e.g., Olson & Fazio, 2006), consumption behavior (e.g., Gibson, 2008), self-esteem (e.g., Dijksterhuis, 2004), disgust (e.g.,Schienle[, Stark, & Vaitl, 2001](https://www.sciencedirect.com/science/article/pii/S0887618508001163%22%20%5Cl%20%22bib40)), phobias (e.g., Merckelbach, et al., 1993) and much more. In the applied domain, it is frequently used as an intervention to address problematic attitudes and behaviors related to addictive substances such as alcohol (e.g., Houben, Schoenmakers, & Wiers, 2010), unhealthy food consumption (e.g., Shaw et al., 2016), and racism (e.g., Lai et al., 2014).

When it comes to theorizing about EC itself, debate is largely led by proponents of dual process (e.g., Gawronski & Bodenhausen, 2006), single process propositional (e.g., De Houwer, 2018), and association formation models (e.g., Jones et al., 2009). Although many variables are used to differentiate between these positions, one has received considerable attention: contingency awareness (e.g., Corneille & Stahl, 2018). Showing that EC effects can occur without contingency awareness is often viewed as supporting dual process and association formation models whereas the opposite is true for propositional models (although see Stahl & Heycke, 2016). So far the general trend of evidence indicates that EC effects are highly dependent on contingency awareness (e.g., Bar-Anan, De Houwer, & Nosek, 2010; Hofmann et al., 2010; Stahl, Unkelbach, & Corneille, 2009). Yet there is one EC paradigm (Olson & Fazio, 2001) that some argue provides evidence for unaware EC effects (e.g., Jones et al., 2010; March et al., 2018).

This task, which is commonly called the ‘surveillance procedure’, consists of a stream of (distractor) stimuli and requires participants to detect and respond to target stimuli. Unbeknownst to them, several of the distractor stimuli are actually CS-US pairs. In this way the task requires people to process the CS-US pairs but directs their attention away from those pairings and towards irrelevant target items (Jones et al., 2010). Following training, self-reported (and implicit) attitudes are assessed. Participants are then asked post-hoc questions to gauge if they noticed the CS-US pairings during the surveillance task. If so, these ‘contingency aware’ participants are excluded from subsequent analyses. If not, then EC shown by ‘contingency unaware’ participants is treated as supporting the idea that EC effects can occur without awareness (e.g., Jones, Olson & Fazio, 2009, 2010; March, Olson & Fazio, 2018).

Since its introduction in 2001, the surveillance task became one of the most frequently cited EC procedures in the literature (over 620 citations in Google Scholar). Several authors have claimed that the surveillance task provides evidences for unaware EC (e.g., March et al., 2018). They then used these effects to forward conceptual arguments on attitudes in general (i.e., that attitudes can emerge even when people are unaware of their origins), and EC in particular (Walther, Nagengast, & Trasselli, 2005). For instance, the implicit misattribution theory of EC is based almost exclusively on the task’s findings (Jones et al., 2009). Still others use this task to change existing attitudes, primarily because of its purported implicit effects (e.g., Choi, & Lee, 2015; Houben et al., 2010; Olson & Fazio, 2006). Yet others argue that the retrospective measures of contingency knowledge used in this work do not reflect ‘unaware’ EC but instead capture recollective memory for CS-US pairings at the time of judgment rather than awareness of CS-US pairings during encoding (e.g., Gawronski & Walther, 2012).

Regardless of whether one subscribes to the awareness or memory position, constructing theories, and using tasks in applied settings, requires strong evidence. We believe that such evidence is currently lacking. Only a handful of published papers (*n* = 10 reporting 23 separate studies) have actually demonstrated EC effects without awareness/recollective memory using the surveillance paradigm. A random-effects meta-analysis of these studies (see https://osf.io/hs32y/) reveals a significant but small effect size, Hedges’ *g* = 0.20, 95% CI [0.13, 0.28]. However, features in the distribution of these effect sizes suggest that this small average effect may actually be inflated by publication or reporting biases. For instance, studies with larger standard errors tend to find larger effect sizes (see Figure 1). Such ‘funnel-plot asymmetry’ usually indicates that null results from small studies may be missing from the literature (Sterne et al., 2011). In addition, a meta-analytic selection model assuming publication bias (Vevea & Hedges, 1995) fit the data better than a standard random-effects meta-analysis, χ2(1) = 6.49, *p* = .011, and reveals a non-significant average effect size, Hedges’ *g* = 0.07, 95% CI [-0.006, 0.14]. It is therefore possible that the available evidence of EC effects generated using the surveillance paradigm is biased by the selective publication of significant results.

In short, the surveillance task is thought to provide evidence for EC effects without awareness/recollective memory, is used to advocate for dual-process and association formation models of EC and attitudes, and is often deployed as an intervention to ‘implicitly’ modify problematic attitudes and behavior. Such developments seem premature given that few studies exist, and those that do are characterized by small samples and very small effect sizes. Given the theoretical and practical implications stemming from this task, it seems prudent to replicate the basic effect with a highly powered sample. Doing so will provide a strong constraint on future theorizing about attitudes, EC, and the use of this task in applied contexts.

Towards this end, we contacted the original authors and asked for their assistance in designing a procedure that directly replicated their original (2001) procedure. Rather than replicate their original design, the original authors recommended that we make some changes to the study design, based on their own experiences with the task, and on the assumption that this would maximize our chances of obtaining an effect (e.g., March et al., 2018). Our final and approved (by the original authors) study protocol can be found at the following link (https://osf.io/hs32y/). The original authors also recommended that we run the experiment locally in the laboratory rather than on-line. In order to do so, and to collect the necessary sample size, we contacted several labs with extensive expertise with EC to help with data collection. Twelve labs (including the lab of one of the original authors) agreed to contribute to this replication effort.

In addition to replicating the original study, we wanted to explore whether evidence for EC in this task depends on the specific way in which contingency awareness/recollective memory is measured. The original authors contingency awareness criteria may have accidentally included individuals who were actually aware of/remembering the contingencies. We therefore added three questions to the end of the original procedure that will allow us to assess contingency awareness/recollective memory in a more conservative manner.

**Disclosures**

All materials, data, analyses and code will be made available on the Open Science Framework (https://osf.io/hs32y/). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data will be collected in accordance with the Declaration of Helsinki. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

**Method**

**Participants**

Participants will be recruited from twelve labs at ten universities throughout Europe and North America. In each case, only native speaking participants will be recruited. Each lab will collect data from at least 100 participants, and a maximum of 150 participants, on the basis of their local resources. In previously published studies the proportion of contingency aware participants ranged from 2% to 27%. Consequently, 1,200 participants will allow for greater than 99% power to observe a small EC effect (Cohen’s *d* = 0.20) even if 30% of the sample are subsequently excluded on the basis of contingency awareness/recollective memory.[[1]](#footnote-2)

**Materials**

**Unconditioned stimuli**. Study materials provided by the original authors will be used. Ten positive words, ten negative words, ten positive images, and ten negative images will serve as the USs. The positive (*Useful, Calming, Desirable, Appealing, Worthwhile, Relaxing, Beneficial, Valuable, Terrific, Commendable*) and negative words (*Inferior, Harmful, Offensive, Troublesome, Upsetting, Terrifying, Unhealthy, Useless, Dislikable, Undesirable*) are identical to those used in Experiment 5 of Jones et al. (2009).[[2]](#footnote-3) The positive and negative images were originally selected from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1997) or the web. However, due to the quality of the original images, we were only able to use nine of the ten positive and nine of the ten negative images from the Jones et al. (2009) study. In consultation with the original authors, we therefore chose two additional IAPS images – one positive and the other negative.

**Conditioned stimuli**. For the conditioned stimuli, the original authors recommended that we not use their original CSs because these items may be relatively familiar to modern samples (see Jones et al., 2009). Instead they advised us to select stimuli that are relatively novel and neutral to the sample population. Based on this recommendation we generated a set of sixty Pokémon characters. We pretested these characters along two dimensions (valence and familiarity) with a separate sample of 155 participants using the Prolific Academic website (https://prolific.ac) (see https://osf.io/hs32y/). On the basis of this pretest we then selected the twenty characters that were rated as most neutral and least familiar. Participating labs will be instructed to further pretest these twenty characters onsite in order to identify the nine characters that are most neutral and least familiar to participants at that specific lab. The two characters that (a) are most neutral and least familiar, and (b) which differ least in valence and familiarity will serve as CSs. Labs unable to carry out such a pretest will use the nine characters derived from our own initial pretest. In this case two characters (Palpitoad and Bergmite) will serve as CSs.

**Filler and target stimuli**. The seven characters not selected during the pre-rating phase to serve as CSs (*see above*) will serve as target and filler stimuli. Finally, six neutral words (*Book*, *Concrete*, *Umbrella*, *Pencils*, *Glasses*, *Computer*) and four neutral IAPS images will serve as filler stimuli. The original authors did not provide us with filler items and therefore we had to select these items and have them approved by those authors.

**Procedure**

 Participants will complete four tasks in fixed order (surveillance task, filler task, evaluation task, post-experimental questionnaire) and will do so in their native language. The assignment of CS to US valence will be counterbalanced between participants.

**Surveillance task.** The surveillance task consists of 5 blocks, each containing a differenttarget stimulus. Each block is comprised of 86 trials, each presented for 1,500ms with no inter-trial interval. Each block will include 8 CS-US pair trials (4 CS-USpos trials and 4 CS-USneg trials), 10 target trials, 30 blank screen trials, and 38 fillers trials. In all cases (except for blank screen trials) one or two stimuli will be presented on-screen. Each CS-US pair will be preceded and followed by a blank screen trial, and these ‘triplets’ will be fixed at various positions throughout the procedure (10-12, 20-22, 30-32, 40-42, 50-52, 60-62, 70-72, 80-82, with an alternation between the CSpos and CSneg). The assignment of CS-US pairs to the fixed positions will occur randomly. As recommended by Jones et al. (2009), the CS and the US will be presented close to one another (approximately 1cm from each other) and the CS will always be larger than the US. In each block, target trials, filler trials, and 14 blank screen trials will be presented randomly in the remaining locations (see https://osf.io/dw2bg/ for a detailed overview of trial content).

Prior to the surveillance task participants will read the following instructions:

*Imagine that you are a security guard watching for deviant activity at a business. Your job requires that you pay attention at all times, and respond quickly when something suspicious happens. In our lab we study attention and rapid responding, and in this experiment you’ll be asked to play the role of the security guard.*

*Specifically, you will be attending to a number of items presented on the computer screen, and you’ll be responding as quickly as possible when a target item appears by pressing the spacebar. The target item will appear at random several times throughout the experiment. The target item may appear as an image or as a name. So be sure to pay attention at all times and focus on the screen, because you never know when the target item will appear. A number of filler items that we’ve selected from our stimulus pool will also be shown randomly to make the task more challenging. These distractors are both pictures and words that were just randomly picked from our collection.*

*Sometimes two images will appear on the screen at the same time, and sometimes only one image will appear. Be sure to hit the spacebar only when the target appears. The target might appear anywhere on the screen as well, and it might also appear with other images. So whenever you see a target image or name anywhere on the screen, hit the spacebar.*

*The items will be displayed rapidly, so make sure that when you see a target, you hit the spacebar before it disappears. Again, be sure to pay close attention throughout the experiment so that you can respond as quickly and accurately as possible.*

*There will be five separate surveillance tasks of about 4 minutes each. Each task will have a different target, and all of the target items will be cartoon creatures.*

**Filler task.** Although a filler task has not been used in the majority of published studies with the surveillance task, the original authors recommended that a filler task be added in order to create a delay between that task and the evaluation task (e.g., Kendrick & Olson, 2012). The filler task will include two questionnaires: the Need for Cognition scale (18-item NFC Scale; Cacioppo, Petty, & Kao, 1984) and the Need to Evaluate scale (16-item NFE scale; Jarvis & Petty, 1996), presented in a fixed order (NFC followed by NFE). These tasks are not central to the main hypotheses being tested here and will not be subsequently analyzed. Nevertheless, those interested in this data can retrieve it from the OSF website (https://osf.io/hs32y/).

**Evaluation task.** Following the filler task, participants will complete a 30 trial forced-choice task (Jones et al., 2009). On each trial, a pair of stimuli will be presented onscreen and participants will indicate as quickly as possible which image they prefer by pressing a corresponding key. Ten of the trials will present one or both CSs (two will present the CSpos and CSneg together, four will present the CSpos with one of the neutral targets/fillers, and four will present CSneg with one of the neutral targets/fillers). The remaining 20 trials will be filler trials, each presenting two neutral targets/fillers. Two filler trials will always precede the first critical trial, and subsequent critical trials will appear at fixed points separated by filler trials (positions 3, 6, 9, 12, 15, 18, 21, 24, 27 and 30). The ten critical trials will be randomly assigned to the fixed positions.

Participants will see the following instructions:

*Next, you’ll be presented with 30 pairs of target and filler creatures from the surveillance tasks, and we’d like you to indicate which one you like better. You don’t need a reason for liking one rather than the other, just give us your gut feelings. We are interested in knowing if the pleasantness or unpleasantness of these stimuli affects the ability to attend and rapidly respond to them, so we need you to indicate which you prefer. Remember, you don’t need a reason for liking one rather than the other, so just go with your gut. Please respond quickly.*

**Post-experimental questionnaire**. After the evaluation task, participants will complete a questionnaire: we will use the original Olson and Fazio (2001) post-experimental questionnaire followed by the questionnaire used in the studies of Bar-Anan et al. (2010). With respect to the former, participants will first answer three open-ended questions: 1. *Think back to the very first part of the experiment. Did you notice anything out of the ordinary in the way the words and pictures were presented during the surveillance tasks?* 2. *Did you notice anything systematic about how particular words and images appeared together during the surveillance tasks?* 3. *Did you notice anything about the words and images that appeared with certain cartoon creatures?*. Although the original authors recommended that we collect data for all three questions, they also recommended that we only use the first two questions when assessing awareness.

With respect to the Bar-Anan et al. (2010) protocol, participants will be asked the following three questions: 1. *For some participants, during the first task, there was one cartoon creature that always appeared with positive images and words, and one that always appeared with negative images and words. Do you think it happened in your case?* (response options: No, I did not notice if that happened in my task, Yes, that happened in my task). 2. *During the first task, which of the two characters was consistently presented with positive images and words?* 3. *During the first task, which of the two characters was consistently presented with negative images and words?* (response options to questions 2 and 3: CSpos (certainly), CSpos (probably), CSpos (guess), CSneg (guess), CSneg (probably), CSneg (certainly). Finally, we will assess for familiarity with the Pokémon presented in the task: *How familiar were you with the cartoon creatures that appeared in the surveillance tasks?* (response scale: 0 = *Not familiar at all* to 8 = *Very familiar*).

**Experimental fidelity.** We have taken a number of steps in order to maximize experimental fidelity across labs. First, given differences in the native languages of participating labs (e.g., Dutch, German, Spanish, French, Polish), materials originally produced in English will be translated. We will do so using a forward and backward translation process. Specifically, materials will first be translated from English into the native language used at a given lab by one member of that participating team. This translation will then be backward translated into English by another member of that same team who was not involved in the initial translation process. This backward translation will be returned to the coordinating team for verification and approval. If necessary (i.e., where the backward translation is not approved) the translation process will be repeated until approval is provided. Second, the entire experimental protocol will be standardized across all labs. Specifically, each lab will run the experiment using the same program and general materials (i.e., developed in PsychoPy; Peirce, [2007](https://www.tandfonline.com/doi/full/10.1080/02699930903485076)) which will generate identically formatted raw data files across all sites. We will then collate these data files from all sites and analyze them centrally using a single set of R code and scripts. All materials and analytic files will be pre-registered before data collection begins (see https://osf.io/hs32y/).

**Planned Analyses**

**Data Preparation**

 **Surveillance task**. We will compute the number of errors made during the surveillance task for each participant (errors are defined as responding to non-target trials, or not responding to target trials). Doing so will allow us to check if participants paid attention during that task. Based on the original authors’ recommendations, we will exclude participants who are more than three standard deviations above or below the mean number of errors.

 **Evaluation task**. Following Jones et al. (2009), a *self-reported preference score* will be calculated for each participant based on their performance during the evaluation task. Specifically, a score of 1 will be assigned to trials in which the participant chooses the CSpos or the image appearing together with CSneg. A score of −1 will be assigned to trials in which participants chose the CSneg or the image appearing together with CSpos. The sum of this coding,which can range from -10 to +10 will serve as measure of evaluative responding (i.e., a preference for CSpos over CSneg).

 **Contingency awareness/recollective memory**.

***Confirmatory analyses*.** We will compute a score following the original authors’ recommendations. Specifically, two independent raters (from each lab) will code participants’ free responses to questions 1-2 from the original authors’ questionnaire and judge whether those responses show correct identification of the CS-US pairings. The coding in all labs will be based on the same protocol (see https://osf.io/hs32y/). As recommended by the original authors, participants will be excluded if both raters agree that participants identified the valence of the USs that were paired with each of the CSs, in at least one of the two questions. If participants identify that one of the CSs was paired with a US of a particular valence, or report that CSs and USs were paired during the task (even if they do not mention the specific way in which they were paired), then they will be retained and coded as being ‘contingency unaware’. Likewise, in cases of rater disagreement, participants will also be retained and coded as ‘contingency unaware’ as per the original authors criteria.

***Exploratory analyses*.** As we outlined in the introduction, the original authors criteria may have accidentally included individuals who were aware of/remembering the contingencies. Therefore we will compute three additional exploratory scores to examine if evidence for EC in this task depends on the specific way in which contingency awareness/recollective memory is measured.

The first (exploratory) score will use a more conservative coding of the original authors’ questions. Participants will be coded as ‘aware’ if they express full or partial memory. Specifically, assignment to the ‘aware’ group will occur when both judges agree that the participant identified the valence of the USs that were paired with each of the CSs, or identified that one of the CSs was paired with a US of a particular valence, or reports that CSs and USs were paired during the task (even if they do not mention the specific way in which they were paired), in at least one of the two questions. Assignment to the ‘unaware’ group will occur when both judges indicate that the participant did not report that CSs were systematically paired with USs, or that a CS was paired with a US of a specific valance, in at least one of the two questions. In cases of rater disagreement, a third judge will be recruited and asked to provide their own judgement according to the above criteria. The majority judgement will be adopted. Participants in the ‘aware’ group will be excluded from subsequent analysis.

 The second (exploratory) score will be computed based on Bar-Anan et al.’s (2010) criteria. Here participants will be excluded if they chose the “yes” answer on question 1 of the Bar-Anan et al. measure, and retained if they chose “no”.

The third (exploratory) score will be computed based on a modification to Bar-Anan et al.’s (2010) criteria in order to verify that responding is not driven by guessing. Participants will be excluded if they chose the “yes” answer on question 1 and correctly identify the valence with which each of the two CSs appeared during the task (providing either a *probably* or *certainly* response on questions 2-3). All other participants will be retained.

In all previous analyses, ‘contingency-aware’ participants will be excluded. Yet one could also examine if awareness/recollective memory moderates the size of EC effects. With this in mind, we will first divide participants into two groups (‘aware’ and ‘unaware’) using the (four) aforementioned criteria, and then carry out an additional set of exploratory analyses that compare EC effects between these two groups. [[3]](#footnote-4)

**Analytic Strategy**

**Confirmatory analyses**. To determine if EC effects emerge in the absence of contingency awareness/recollective memory, according to the original authors criteria, we will compute the EC effect size (Hedges’ g) from the mean and standard deviation of the self-reported preference score in the ‘unaware’ group. Thereafter we will meta-analyze these effect sizes in a meta-analysis using a random-effects model, using an alpha value of 0.05. Although all participating labs will use similar materials, differences may be introduced by the translation of materials, selection of stimuli, or characteristics of the samples. In order to account for this within the analyses, we will employ random effects meta-analysis models (specifically, using the Restricted Maximum Likelihood method).

The meta-analysis based on the original authors’ criteria showed that, on average, the surveillance task led to a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX, in the [expected/unexpected] direction. The EC effect size in this group ranged from X.XX to X.XX across labs (see Figure X). The differences in EC effect sizes across labs were [consistent/inconsistent] with what one would expect by chance, τ = X.XX, *I*2 = X.XX%, *H*2 = X.XX, Q(X) = X.XX, *p* = .XXX.

**Confirmatory hypotheses**. Based on the above analyses, these findings [replicate/do not replicate] the original authors findings.

**Exploratory analyses**.

***EC effects in the absence of contingency awareness/recollective memory*.** Three different groups will be created (i.e., those based on the modification to the original authors’ criteria, those based on the original Bar-Anan et al., criteria, and those based on the modified Bar-Anan et al. criteria). For each group (in each lab) we will compute the EC effect size (Hedges’ g) from the mean and standard deviation of the self-reported preference score. Thereafter we will meta-analyze these effect sizes in three independent meta-analyses using a random-effects model.

The meta-analysis with the first exploratory criteria (i.e., the modified original authors’ criteria) showed that, on average, the surveillance task led to a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX, in the [expected/unexpected] direction. The EC effect size in this group ranged from X.XX to X.XX across labs (see Figure X). The differences in EC effect sizes across labs were [consistent/inconsistent] with what one would expect by chance, τ = X.XX, *I*2 = X.XX%, *H*2 = X.XX, Q(X) = X.XX, *p* = .XXX.

The meta-analysis with the second exploratory criteria (i.e., the original Bar-Anan et al. criteria) showed that, on average, the surveillance task led to a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX, in the [expected/unexpected] direction. The EC effect size in this group ranged from X.XX to X.XX across labs (see Figure X). The differences in the EC effect size across labs were [consistent/inconsistent] with what one would expect by chance, τ = X.XX, *I*2 = X.XX%, *H*2 = X.XX, Q(X) = X.XX, p = .XXX.

The meta-analysis with the third exploratory criteria (i.e., the modified Bar-Anan et al. criteria) showed that, on average, the surveillance task led to a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX, in the [expected/unexpected] direction. The EC effect size in this group ranged from X.XX to X.XX across labs (see Figure X). The differences in the EC effect size across labs were [consistent/inconsistent] with what one would expect by chance, τ = X.XX, *I*2 = X.XX%, *H*2 = X.XX, Q(X) = X.XX, p = .XXX.

Finally, to investigate if the effect sizes computed based on the four awareness/recollective memory criteria differ from one another, we used a multilevel meta-analysis with the type of criteria as a moderator, adding a random intercept for laboratory to account for the statistical dependency between effect sizes coming from related samples. The moderator test showed that the results of the four criteria [differed/did not differ] from each other, *Q*(X) = X.XX, *p* = .XXX.

***Comparison of ‘Contingency-Aware’ vs. ‘Unaware’ Participants.*** We meta-analyzed EC effect sizes for ‘contingency-aware’ participants and compared these to the effect sizes obtained from ‘contingency-unaware’ participants via moderator analyses. We did so in order to examine if the ‘contingency-aware’ participants excluded in previous analyses produced higher or lower EC effects than their ‘contingency-unaware’ counterparts. All moderator analyses reported in this section included a random intercept at the laboratory level in order to account for the dependencies between effect sizes coming from the same experimental setting.

First, participants classified as ‘aware’ according to the original authors’ criteria showed a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX. The moderator test showed that this effect size [differed/did not differ] from the effect observed in unaware participants, *Q*(X) = X.XX, *p* = .XXX.

Second, participants classified as ‘aware’ according to the modified original authors criteria showed a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX. The moderator test that this effect size [differed/did not differ] from the effect observed in unaware participants, *Q*(X) = X.XX, *p* = .XXX.

Third, participants classified as ‘aware’ according to the original Bar-Anan et al. criteria showed a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX. The moderator test that this effect size [differed/did not differ] from the effect observed in unaware participants, *Q*(X) = X.XX, *p* = .XXX.

Finally, participants classified as ‘aware’ according to the modified Bar-Anan et al. criteria showed a [significant/non-significant] and [small/medium/large] EC effect size, Hedges’ *g* = X.XX, 95% CI [X.XX, X.XX], *z* = X.XX, *p* = .XXX. The moderator test that this effect size [differed/did not differ] from the effect observed in unaware participants, *Q*(X) = X.XX, *p* = .XXX.

 **Exploratory hypotheses**.

***EC effects in the absence of contingency awareness/recollective memory.*** There are three outcomes that we have *a priori* hypotheses for. The first is a situation where the multilevel meta-analysis returns a significant overall EC effect, but no significant effect for the type of criteria. In this case we will conclude that EC effects do emerge in the surveillance task and do not depend on the specific way in which contingency awareness/recollective memory is measured. The second is where we find no evidence for an overall EC effect and the type of criteria also fails to moderate the size of EC. In this case we will conclude that EC effects do not emerge in the surveillance task. The third is where we find a significant effect of type of criteria in the multilevel meta-analysis and the individual univariate meta-analysis reveal significant evidence for EC with the original authors criteria but with none of the other three criteria. In this case we will conclude that EC effects in the surveillance task strongly depend on the way that the original authors chose to assess contingency awareness/recollective memory.

***Comparison of ‘Contingency-Aware’ vs. ‘Unaware’ Participants*.** We hypothesize that EC effects will be larger for contingency-aware than for contingency-unaware participants, although as mentioned in Footnote 3, the results of this analysis must be interpreted with caution.

 **Authorship**

TM contributed to project administration, the creation of the procedure protocol, the design of the materials, writing the original draft, and review and editing the final manuscript. SH contributed to project administration, writing the original draft, and review and editing the final manuscript. IH contributed to project administration, the design of the materials and analysis script, and review the manuscript. MAV contributed to writing the original draft, and review and editing the final manuscript. MAO contributed to the creation of the procedure protocol, and review of the manuscript. FA, KB, RB, TB, OC, SBD, MJF, KAF, AG, BG, TH, FH, MH, BK, AM, JR, JS, CTS, CS, PT and CU contributed to the review of the manuscript. JDH contributed to project administration, the creation of the procedure protocol, and review of the manuscript. Authorship order will be updated according to contribution during the project.

**References**

Bar-Anan, Y., De Houwer, J., & Nosek, B. A. (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *The Quarterly Journal of Experimental Psychology, 63*(12), 2313-2335.‏

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.‏

Choi, Y. J., & Lee, J. H. (2015). Alcohol-related attitudes of heavy drinkers: Effects of arousal and valence in evaluative conditioning. *Social Behavior and Personality: an International Journal, 43*(2), 205-215.‏

Corneille, O., & Stahl, C. (2018). Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models. *Personality and Social Psychology Review*. [doi.org/10.1177/1088868318763261](https://doi.org/10.1177/1088868318763261)

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13(3)*, e28046. https://doi.org/10.5964/spb.v13i3.28046

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

Dijksterhuis, A. P. (2004). I like myself but I don't know why: Enhancing implicit self-esteem by subliminal evaluative conditioning. *Journal of Personality and Social Psychology, 86*(2), 345-355.‏

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

Gawronski, B., & Walther, E. (2012). What do memory data tell us about the role of contingency awareness in evaluative conditioning?. *Journal of Experimental Social Psychology*, *48*(3), 617-623.‏

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35*(1), 178-188.‏

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

Houben, K., Schoenmakers, T. M., & Wiers, R. W. (2010). I didn't feel like drinking but I don't know why: The effects of evaluative conditioning on alcohol-related attitudes, craving and behavior. *Addictive Behaviors, 35*(12), 1161-1163.‏

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70*(1), 172-194.‏

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of Personality and Social Psychology, 96*(5), 933-948.‏

Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The “how” question. *Advances in Experimental Social Psychology, 43*, 205–255.

Kendrick, R. V., & Olson, M. A. (2012). When feeling right leads to being right in the reporting of implicitly-formed attitudes, or how I learned to stop worrying and trust my gut. *Journal of Experimental Social Psychology, 48*(6), 1316-1321.‏

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., ... & Frazier, R. S. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765-1785.‏

Lang, P. J., Bradley, M. M. & Cuthbert, B. N. (1995). International Aective Picture System: Technical manual and affective ratings. Gainesville, FL: University of Florida

March, D. S., Olson, M. A., & Fazio, R. H. (2018). The implicit misattribution model of evaluative conditioning. *Social Psychological Bulletin, 13*, e27574.‏

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413-417.‏

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421-433.‏

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.‏

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, Vol. 19 (pp. 123–205). New York: Academic.

Schienle, A., Stark, R., & Vaitl, D. (2001). Evaluative conditioning: A possible explanation for the acquisition of disgust responses?. *Learning and Motivation, 32*(1), 65-83.‏

Shaw, J. A., Forman, E. M., Espel, H. M., Butryn, M. L., Herbert, J. D., Lowe, M. R., & Nederkoorn, C. (2016). Can evaluative conditioning decrease soft drink consumption?. *Appetite, 105*, 60-70.‏

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

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

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., ... & Tetzlaff, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj, 343*, d4002.‏

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (in press). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General.*

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435.‏

Walther, E., Nagengast, B., & Trasselli, C. (2005). Evaluative conditioning in social psychology: Facts and speculations. *Cognition and Emotion, 19*(2), 175-196.‏

****

**Figure 1.** Funnel plot of the data entered into the meta-analysis of previous studies with the surveillance task. Each dot depicts effect size (Hedges’ *g*) against the SE. Studies falling inside the grey area are statistically non-significant in a two-tailed test. The triangle inside the dashed line is centered at the average mean effect size and represents the distribution of effect sizes that would be expected in the absence of publication bias. The red line represents Egger’s regression test for funnel plot asymmetry.

1. The minimum sample size has 99% power to detect Cohen’s d of 0.13 and 80% power to detect Cohen’s d of 0.08 (within subjects, one tailed, alpha = 0.05). [↑](#footnote-ref-2)
2. The original authors also recommended that we use mildly evocative stimuli in this experiment. [↑](#footnote-ref-3)
3. Note that the results obtained from such a comparison should be interpreted with extreme caution. First, any attempt to detect differences in EC effects between putatively ‘aware’ and ‘unaware’ participants will ultimately depend on the reliability of the awareness measure used, and of the EC procedure itself. Previous evidence suggests that unconscious learning paradigms and awareness tests tend to yield unreliable measures (e.g., Vadillo, Linssen, Orgaz, Parsons, & Shanks, in press). Second, it is conceptually and statistically problematic to use one outcome measure as a moderator of another outcome measure, due to the correlational nature of their relation (e.g., Gawronski & Walther, 2012). We will unpack both issues in greater detail in the General Discussion. [↑](#footnote-ref-4)