











Incidental Attitude Formation via the Surveillance Task: A Preregistered Replication of the Olson and Fazio (2001) Study



Tal Moran¹, Sean Hughes¹, Ian Hussey¹, Miguel A. Vadillo², Michael A. Olson³, Frederik Aust⁴, Karoline Bading⁵, Robert Balas⁶, Taylor Benedict⁴, Olivier Corneille⁷, Samantha B. Douglas⁸, Melissa J. Ferguson⁹, Katherine A. Fritzen³, Anne Gast⁴, Bertram Gawronski¹⁰, Tamara Giménez-Fernández², Krzysztof Hanusz⁶, Tobias Heycke¹¹, Fabia Högden⁴, Mandy Hütter¹², Benedek Kurdi⁹, Adrien Mierop⁷, Jasmin Richter⁴, Justyna Sarzyńska-Wawer⁶, Colin Tucker Smith⁸, Christoph Stahl⁴, Philine Thomasius⁴, Christian Unkelbach⁴, and Jan De Houwer¹

¹Department of Experimental Clinical and Health Psychology, Ghent University; ²Department of Psychology, Universidad Autónoma de Madrid; ³Department of Psychology, University of Tennessee; ⁴Department of Psychology, University of Cologne; ⁵Institute of Psychology, Friedrich Schiller University Jena; ⁶Institute of Psychology, Polish Academy of Sciences; ⁷Psychological Sciences Research Institute, Université Catholique de Louvain; ⁸Department of Psychology, University of Florida; ⁹Department of Psychology, Cornell University; ¹⁰Department of Psychology, University of Texas at Austin; ¹¹Department Survey Design & Methodology, GESIS–Leibniz Institute for the Social Sciences; and ¹²Department of Psychology, Eberhard Karls Universität Tübingen

Abstract

Evaluative conditioning is one of the most widely studied procedures for establishing and changing attitudes. The surveillance task is a highly cited evaluative-conditioning paradigm and one that is claimed to generate attitudes without awareness. The potential for evaluative-conditioning effects to occur without awareness continues to fuel conceptual, theoretical, and applied developments. Yet few published studies have used this task, and most are characterized by small samples and small effect sizes. We conducted a high-powered ($N = 1,478$ adult participants), preregistered close replication of the original surveillance-task study (Olson & Fazio, 2001). We obtained evidence for a small evaluative-conditioning effect when “aware” participants were excluded using the original criterion—therefore replicating the original effect. However, no such effect emerged when three other awareness criteria were used. We suggest that there is a need for caution when using evidence from the surveillance-task effect to make theoretical and practical claims about “unaware” evaluative-conditioning effects.

Keywords

preregistered replication, evaluative conditioning, contingency awareness, recollective memory, attitude formation, open data, open materials, preregistered

Received 3/27/19; Revision accepted 7/9/20

Evaluative conditioning is a widely studied and highly applicable procedure for establishing and changing attitudes (e.g., De Houwer, Thomas, & Baeyens, 2001). In a typical evaluative-conditioning task, a neutral

Corresponding Author:

Tal Moran, Ghent University, Department of Experimental Clinical and Health Psychology
E-mail: tmo286@gmail.com

Psychological Science
1–12

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0956797620968526

www.psychologicalscience.org/PS



stimulus (or *conditioned* stimulus [CS]) is repeatedly paired with a positive or a negative stimulus (or *unconditioned* stimulus [US]), and as a result, the former acquires a similar valence to the latter.

Evaluative conditioning plays a central role in theory and application throughout psychological science. For instance, the associative–propositional evaluation model (Gawronski & Bodenhausen, 2006), an influential theory of attitudes in social psychology, distinguishes between explicit attitudes and implicit attitudes and treats evaluative conditioning as a key pathway for changing the latter. The elaboration-likelihood model, in the domain of persuasion (Petty & Cacioppo, 1986), distinguishes between the central and peripheral routes to persuasion and views evaluative conditioning as highly relevant to the latter route. Elsewhere, evaluative conditioning 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), phobias (e.g., Merckelbach, de Jong, Arntz, & Schouten, 1993), and many other domains. 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 evaluative conditioning, debate is largely led by proponents of dual-process models (e.g., Gawronski & Bodenhausen, 2006), single-process propositional models (e.g., De Houwer, 2018), and associative models (e.g., Jones, Fazio, & Olson, 2009). Although many variables are used to differentiate between these positions, one has received considerable attention: contingency awareness (e.g., Corneille & Stahl, 2019). Showing that evaluative-conditioning effects can occur without contingency awareness is often viewed as support for dual-process and associative models, whereas the opposite is true for propositional models (although see Stahl & Heycke, 2016). So far, the general trend of evidence indicates that evaluative-conditioning effects are highly dependent on contingency awareness (e.g., Bar-Anan, De Houwer, & Nosek, 2010; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; Stahl, Unkelbach, & Corneille, 2009). Yet there is one evaluative-conditioning paradigm (Olson & Fazio, 2001) that some researchers argue provides evidence for “unaware” evaluative-conditioning effects (e.g., Jones, Olson, & Fazio, 2010; March, Olson, & Fazio, 2018).

In this task, commonly called the “surveillance task,” neutral and valenced stimuli are surreptitiously paired while the participants complete an unrelated task. Two neutral and unfamiliar Pokémon are selected to serve as CSs. Valenced pictures and words serve as USs.

Statement of Relevance

How can we influence people’s attitudes? A powerful method is evaluative conditioning. In evaluative conditioning, a neutral stimulus (e.g., a new product) is paired with either a positive or a negative stimulus (e.g., puppies). The pairing can change how people feel toward the first (neutral) stimulus. A widespread assumption is that evaluative conditioning can change attitudes even when people are unaware that stimuli are being paired. We tested this idea with more than 1,400 participants using the surveillance task, a procedure that purportedly generates attitudes in the absence of awareness. We found that new attitudes were formed in the absence of awareness only when we used a specific, narrow definition of awareness. When awareness was defined more broadly, attitudes did not change outside of awareness. These findings suggest caution when using evidence from the surveillance task to construct theories about how attitudes are formed and to design interventions to implicitly modify problematic beliefs and behavior.

Participants are told that they will take part in a surveillance task wherein they have to detect several target Pokémon that are different from the actual Pokémon of interest (i.e., the CSs) and press a key when they see them. During the task, participants encounter many trials, some of which present a target Pokémon to which they have to respond, and others present (distractor) stimuli to which they do not need to respond. Unbeknownst to them, several of the distractor trials present CS-US pairs. Specifically, on some of the distractor trials, one Pokémon (CS1) is always presented alongside a positive word or image (positive US), whereas on other distractor trials, a second Pokémon (CS2) is always presented with a negative word or image (negative US). In this way, the task requires people to process the CS-US pairs but directs their attention away from those pairings and toward the irrelevant target stimuli. Afterward, relative preferences for CS1 and CS2 are assessed, followed by retrospective measures of awareness of the CS-US contingencies that were present during the surveillance task. Researchers who use this task assume that people will prefer CS1 (i.e., the Pokémon paired with positive stimuli) over CS2 (i.e., the Pokémon paired with negative stimuli), even if they later report no awareness of the CS-US contingencies (e.g., Jones et al., 2009; Jones et al., 2010; March et al., 2018).

Since its introduction in 2001, the surveillance task has become one of the most frequently cited evaluative-conditioning procedures in the literature (more than 700 citations in Google Scholar as of June 2020). Several

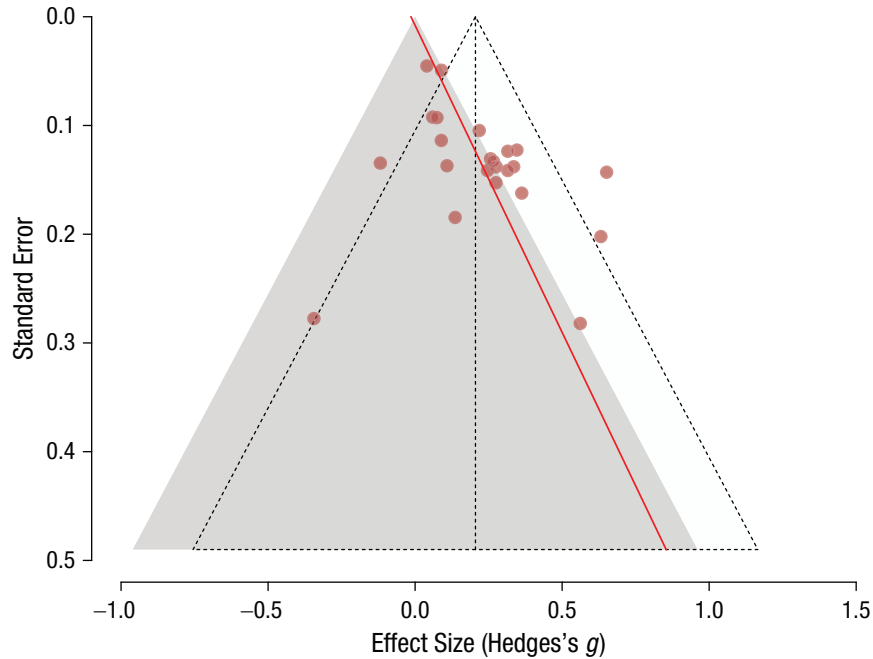


Fig. 1. Funnel plot of the data entered into the meta-analysis of previous studies that used the surveillance task. Each dot depicts the effect size (Hedges's g) from a single study as a function of its standard error. Studies falling inside the gray area were statistically nonsignificant in a two-tailed 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.

authors have claimed that the surveillance task provides evidence for unaware evaluative conditioning (e.g., March et al., 2018). They then used these effects to promote conceptual arguments on attitudes, in general (i.e., that attitudes can emerge even when people are unaware of their origins), and evaluative conditioning, in particular (Walther, Nagengast, & Trasselli, 2005). For instance, the implicit misattribution theory of evaluative conditioning is based almost exclusively on the task's findings (Jones et al., 2009). Still other authors have used 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). Other researchers have argued that the retrospective measures of contingency knowledge used in this work do not reflect unaware evaluative conditioning 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, strong evidence is required to construct theories and use tasks in applied settings. We believe that such evidence is currently lacking. Only a handful of published articles ($n = 10$ reporting 23

separate studies) have supported the possibility of evaluative-conditioning effects without awareness/recollective memory using the surveillance paradigm. A random-effects meta-analysis of these studies (see <https://osf.io/4mh2d>) revealed a significant but small effect size (Hedges's $g = 0.20$, 95% confidence interval, or CI = [0.13, 0.28]). However, features in the distribution of these effect sizes suggest that this small average effect may be inflated by publication or reporting biases. For instance, studies with larger standard errors tend to find larger effect sizes (see Fig. 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) fitted the data better than a standard random-effects meta-analysis, $\chi^2(1) = 6.49$, $p = .011$, and revealed a nonsignificant average effect size (Hedges's $g = 0.07$, 95% CI = [-0.006, 0.14]). It is therefore possible that the available evidence of evaluative-conditioning effects generated using the surveillance paradigm is biased by the selective publication of significant results.

In short, some researchers argue that the surveillance task provides evidence for evaluative-conditioning

effects without awareness/recollective memory, is used to advocate for dual-process and associative models of evaluative conditioning and attitudes, and is often deployed as an intervention to implicitly modify problematic attitudes and behavior. Such developments seem premature given that few existing studies support the possibility of evaluative-conditioning effects without awareness/recollective memory, 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 seemed prudent to replicate the basic effect with a highly powered sample. Doing so would provide a strong constraint on future theorizing about attitudes, evaluative conditioning, and the task's use in applied contexts.

Toward this end, we contacted Olson and Fazio and asked for their assistance in designing a procedure that directly replicated their original procedure (Olson & Fazio, 2001). They encouraged us to make changes to their original study design rather than directly replicate it, on the basis of 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). It is therefore important to note that the present study was a close conceptual replication rather than a direct replication of the Olson and Fazio (2001) study. The final study protocol was approved by the original authors (see <https://osf.io/wnckg>). Olson and Fazio also recommended that we run the experiment locally in the laboratory rather than online. To do so, and to collect the necessary sample size, we contacted several labs with extensive expertise in evaluative conditioning to help with data collection. Twelve labs, including the lab of one of the original authors (M. A. Olson), agreed to contribute to this replication effort.

In addition to replicating the original study, we wanted to explore whether evidence for evaluative conditioning in this task depends on the specific way in which contingency awareness/recollective memory is measured. Olson and Fazio's contingency-awareness criterion may have inadvertently included individuals who were aware of or remembering the contingencies. We therefore included three additional measures of contingency awareness/recollective memory that assessed this construct in a more conservative manner.

Method

Participants

We recruited 1,478 adult participants from 12 labs at 10 universities in Europe and North America (72% women, 27% men, < 1% other identity; age: $M = 21.2$ years,

$SD = 4.9$). All labs used an ad hoc sampling strategy to sample from undergraduate students, and all experimental sessions were run in person (rather than online). We initially planned that each lab would collect data from a minimum of 100 participants and a maximum of 150 participants on the basis of their local resources. The rationale for this planned sample size was that in previously published studies, the percentage of contingency-aware participants ranged from 2 to 27. Consequently, 1,200 participants would allow for greater than 99% power to observe a small evaluative-conditioning effect (Cohen's $d = 0.20$), even if 30% of the sample were subsequently excluded on the basis of contingency awareness/recollective memory.¹ For details on the sample size and characteristics for each lab, see the Supplemental Material available online. All data from all sites were included in the analyses, following the amended preregistration for our data-collection stopping rule (<https://osf.io/uyng7>). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data were collected in accordance with the Declaration of Helsinki.

Materials

Unconditioned stimuli. Study materials provided by Olson and Fazio were used. Ten positive words, 10 negative words, 10 positive images, and 10 negative images served as the USs. The positive words (*useful, calming, desirable, appealing, worthwhile, relaxing, beneficial, valuable, terrific, commendable*) and negative words (*inferior, harmful, offensive, troublesome, upsetting, terrifying, unhealthy, useless, disliked, undesirable*) were identical to those used by Jones et al. (2009) in their Experiment 5.² The positive and negative images were originally selected from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1995) or the Internet. However, because of the quality of the original images, we were able to use only nine of the 10 positive images and nine of the 10 negative images from the Jones et al. (2009) study. In consultation with Olson and Fazio, we therefore chose two additional IAPS images—one positive and the other negative.

Conditioned stimuli. Olson and Fazio recommended that we not use the CSs from their original study because these items may be relatively familiar to modern samples (see Jones et al., 2009). Instead, they advised us to select stimuli that would be relatively novel and neutral to the sample population. On the basis of this recommendation, we generated a set of 20 Pokémon that were pretested in each lab along two dimensions (valence and familiarity). The two characters that (a) were most neutral and least

familiar and (b) differed least in valence and familiarity served as CSs (for more details, see the Supplemental Material; for the results of the pretest conducted at each lab, see <https://osf.io/a3qj9>).

Filler and target stimuli. The seven characters not selected during the pretesting phase to serve as CSs (see above) served as target and filler stimuli. Finally, six neutral words (*book, concrete, umbrella, pencils, glasses, computer*) and four neutral IAPS images served as filler stimuli. Olson and Fazio did not provide us with filler items; therefore, we had to select these items and have them approved by Olson and Fazio.

Procedure

Participants completed four tasks in fixed order (surveillance task, filler task, evaluation task, postexperiment questionnaire) in the lab's native language (for a screen-capture video of the experiment in English, see <https://osf.io/6n4fv>). The assignment of CS to US valence was counterbalanced among participants. Each CS appeared once with each of the 20 USs of the same valence.

Surveillance task. The surveillance task consisted of five blocks, each containing a different target stimulus. Each block comprised 86 trials, each presented for 1,500 ms with no intertrial interval. Each block included eight CS-US pair trials (four CS-positive US trials and four CS-negative US trials), 10 target trials, 30 blank-screen trials, and 38 filler trials. In all cases (except for blank-screen trials), one or two stimuli were presented on screen. Each CS-US pair was preceded and followed by a blank-screen trial, and these "triplets" were fixed at various positions throughout the procedure (10–12, 20–22, 30–32, 40–42, 50–52, 60–62, 70–72, and 80–82, with an alternation between the positive and negative CSs). The assignment of CS-US pairs to the fixed positions occurred randomly. As recommended by Jones et al. (2009), the CS and the US were presented close to one another (~1 cm from each other), and the CS was always larger than the US. In each block, target trials, filler trials, and 14 blank-screen trials were presented randomly in the remaining locations (for a detailed overview of trial content, see <https://osf.io/wnckg>). Prior to completing the surveillance task, participants were instructed to detect the target stimulus and hit the space bar every time a target stimulus appeared (for the specific instructions, see <https://osf.io/wnckg>).

Filler task. Although a filler task was not used in the original study or in the vast majority of published surveillance-task studies (four of the 23 studies in our meta-analysis), Olson and Fazio recommended that we add a filler task to

create a delay between the surveillance task and the evaluation task (e.g., Kendrick & Olson, 2012). The filler task included two questionnaires: the 18-item Need for Cognition (NFC) scale (Cacioppo, Petty, & Feng Kao, 1984) and the 16-item Need to Evaluate (NFE) scale (Jarvis & Petty, 1996), presented in a fixed order (NFC followed by NFE). These tasks were not central to the main hypotheses and were therefore not analyzed. Nevertheless, interested readers can retrieve these data from OSF (<https://osf.io/k9nrf>).

Evaluation task. Following the filler task, participants completed a 30-trial forced-choice task (Jones et al., 2009). On each trial, a pair of stimuli was presented on screen, and participants indicated as quickly as possible which image they preferred by pressing a corresponding key. Ten of the trials presented one or both CSs (two presented the positive and negative CSs together, four presented the positive CS with one of the neutral targets or fillers, and four presented the negative CS with one of the neutral targets or fillers).³ The remaining 20 trials were filler trials, each presenting two neutral targets or fillers. Two filler trials always preceded the first critical trial, and subsequent critical trials appeared at fixed points separated by filler trials (Positions 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30). The 10 critical trials were randomly assigned to the fixed positions (for the instructions preceding the evaluation task, see <https://osf.io/wnckg>).

Postexperiment questionnaire. After the evaluation task, participants completed a questionnaire; we used the original Olson and Fazio (2001) postexperiment questionnaire followed by the questionnaire used in the studies by Bar-Anan et al. (2010). With respect to the former, participants first answered three open-ended questions: (a) "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?" (b) "Did you notice anything systematic about how particular words and images appeared together during the surveillance tasks?" and (c) "Did you notice anything about the words and images that appeared with certain cartoon creatures?" Although Olson and Fazio recommended that we collect data for all three questions, they also recommended that we use only the first two questions when assessing awareness.

With respect to the Bar-Anan et al. (2010) protocol, participants were asked the following three questions: (a) "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”), (b) “During the first task, which of the two characters was consistently presented with positive images and words?” and (c) “During the first task, which of the two characters was consistently presented with negative images and words?” Response options to Questions 2 and 3 were positive CS (certainly), positive CS (probably), positive CS (guess), negative CS (guess), negative CS (probably), and negative CS (certainly). Finally, we assessed 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 took a number of steps to maximize experimental fidelity across labs. First, materials originally produced in English were translated using a forward and backward translation process. Second, the entire experimental protocol was standardized across all labs. Specifically, each lab ran the experiment using the same program and general materials (i.e., developed in PsychoPy; Peirce, 2007), which generated identically formatted raw data files across all sites. We then collated these data files from all sites and analyzed them centrally using a single set of R code (Version 3.6.2; R Core Team, 2019) and scripts.

Results

Data processing

Surveillance task. We computed the number of errors made during the surveillance task for each participant (errors are defined as responding to nontarget trials or not responding to target trials) to check whether participants paid attention during that task. On the basis of Olson and Fazio’s recommendations, we excluded participants who were more than 3 standard deviations above or below the mean number of errors, as in their original study. Two percent of participants were excluded on this basis.

Evaluation task. Following Jones et al. (2009), we calculated a self-reported preference score for each participant on the basis of his or her performance during the evaluation task. Specifically, a score of 1 was assigned to trials in which the participant chose the positive CS or the image that appeared together with the negative CS. A score of –1 was assigned to trials in which participants chose the negative CS or the image appearing together with the positive CS. The sum of this coding, which ranged from –10 to +10, served as a measure of evaluative responding (i.e., a preference for positive CS over negative CS).

Awareness/recollection-memory criteria. Four methods of excluding individuals on the basis of their responses

to the postexperimental questions were preregistered. The first was similar to that employed by Olson and Fazio in their study (Olson & Fazio, 2001), whereas the other three were included to explore the robustness of the effect. These latter criteria either had been used in previously published work (Bar-Anan et al., 2010) or were created by us to provide different levels of stringency around awareness from those previously employed (i.e., higher than the Olson & Fazio, 2001, study and lower than the Bar-Anan et al., 2010, study).⁴

Primary criterion: Olson and Fazio (2001). A score was computed following Olson and Fazio’s recommendations. This score was based on participants’ open-ended responses to the original Olson and Fazio (2001) Postexperimental Questions 1 and 2 (for more details, see the Supplemental Material). Two independent raters, who were blind to one another’s ratings, evaluated responses to these two questions and treated responses on both questions as one (compound) text response (for the exact coding instructions provided to the data-collection sites, see <https://osf.io/2dm6u>). Participants were scored as “aware” if their responses to either of the two questions made correct reference to both of the CS-US pairings. If they failed to meet this criterion for any reason, then they were scored as “unaware.” Scores were then compared between raters so that each participant could be assigned a single score. Participants were scored as aware only if both raters scored them as aware.

Secondary criteria. Olson and Fazio’s criterion may have led individuals who were aware to be scored as if they were unaware. We therefore preregistered three additional exclusion criteria to examine whether evidence for evaluative-conditioning effects in this task was robust to, or depended on, the specific way in which contingency awareness/recollective memory were measured. As detailed in the Supplemental Material, the three alternative exclusion rules categorized participants as aware if they (a) referred to any form of systematic pairing between the CS and US stimuli (Olson & Fazio, 2001, modified criterion); (b) indicated that one CS was systematically paired with positive USs and a second CS was paired with negative USs (Bar-Anan et al., 2010, criterion); or (c) in addition to (b) also correctly identified the valence of the USs with which each of the two CSs appeared (Bar-Anan et al., 2010, modified criterion). Compared with Olson and Fazio’s original criteria, these awareness criteria categorized a larger percentage of participants as aware of the CS-US contingency.

Preregistered analyses

In each analysis, to determine whether evaluative-conditioning effects emerged in the absence of contingency

awareness/recollective memory, we first excluded participants who were scored as aware according to an awareness exclusion criterion and then computed an evaluative-conditioning effect size (Hedges's g) for each site from the mean and standard deviation of the self-reported preference score. Thereafter, we meta-analyzed these effect sizes using an alpha of .05 (two tailed). Although all labs used similar materials, they may have nevertheless differed in the translation of materials, selection of stimuli, or characteristics of the samples. To account for this within the analyses, we employed random-effects meta-analysis models. All analyses were conducted using the R package *metafor* (Viechtbauer, 2010) and used restricted maximum likelihood estimation.

Evaluative-conditioning effects in the absence of contingency awareness/recollective memory.

Primary analyses. The meta-analysis based on the original Olson and Fazio (2001) awareness criterion ($n = 1,340$, 9.2% excluded) showed that, on average, the surveillance task led to a small but significant evaluative-conditioning effect (Hedges's $g = 0.12$, 95% CI = [0.05, 0.20], $z = 3.17$, $p = .002$) in the expected direction. Effect sizes ranged from -0.02 to 0.31 across labs (see Fig. 2a). Variation in effect sizes across sites was consistent with what one would expect by chance (i.e., because of sampling variation alone), $\tau^2 = 0.0$, $I^2 = 0.0\%$, $H^2 = 1.0$, $Q(11) = 5.83$, $p = .885$. In sum, when Olson and Fazio's awareness exclusion criterion was employed, their original effect was replicated.

Secondary analyses. When a modified version of Olson and Fazio's exclusion criterion was applied (Olson & Fazio, 2001, modified; $n = 1,007$, 31.9% excluded), the surveillance task did not produce an evaluative-conditioning effect (Hedges's $g = 0.05$, 95% CI = $[-0.04, 0.13]$, $z = 1.04$, $p = .299$). Effect sizes ranged from -0.08 to 0.30 across sites (see Fig. 2b). Variation in effect sizes across sites was consistent with what one would expect by chance, $\tau^2 = 0.0$, $I^2 = 0.0\%$, $H^2 = 1.0$, $Q(11) = 2.76$, $p = .994$.

When the Bar-Anan et al. (2010) exclusion criterion was applied ($n = 755$, 48.9% excluded), the surveillance task did not lead to an evaluative-conditioning effect (Hedges's $g = 0.03$, 95% CI = $[-0.06, 0.13]$, $z = 0.69$, $p = .493$). Effect sizes ranged from -0.24 to 0.18 across sites (see Fig. 2c). Variation in effect sizes across sites was consistent with what one would expect by chance, $\tau^2 = 0.0$, $I^2 = 0.0\%$, $H^2 = 1.0$, $Q(11) = 4.17$, $p = .965$.

When the modified Bar-Anan et al. (2010) criterion was applied ($n = 1,060$, 28.3% excluded), the surveillance task did not lead to an evaluative-conditioning effect (Hedges's $g = 0.05$, 95% CI = $[-0.03, 0.13]$, $z = 1.17$, $p = .241$). Effect sizes ranged from -0.16 to 0.19 across sites

(see Fig. 2d). Variation in effect sizes across sites was consistent with what one would expect by chance, $\tau^2 = 0.0$, $I^2 = 0.0\%$, $H^2 = 1.0$, $Q(11) = 3.45$, $p = .983$.

Finally, to investigate whether the effect sizes computed on the basis of the four awareness/recollective-memory criteria differed from one another, we combined the data sets used in all of the above analyses into one and used a multilevel meta-analysis with the awareness exclusion criterion as a moderator. A random intercept for data-collection site was included to account for the statistical dependency among effect sizes coming from related samples. The moderator test did not demonstrate evidence that the results of the four criteria differed from each other, $Q(3) = 2.76$, $p = .430$.⁵

Comparison of contingency-aware versus contingency-unaware participants.

The previous analyses excluded contingency-aware participants. Yet one could also examine whether awareness/recollective memory moderates the size of evaluative-conditioning effects. With this in mind, we divided participants into two groups (aware and unaware) using the four aforementioned criteria and then carried out an additional set of secondary analyses that compared evaluative-conditioning effects between these two groups using a multilevel moderator meta-analysis model (for more details about these analyses, see the Supplemental Material). All moderator analyses reported in this section included a random intercept for data-collection site to account for the dependencies between effect sizes coming from the same experimental setting. In each case, we report only the difference between the two conditions (i.e., moderation test) and the effect size in the aware group (effect sizes in the unaware groups can be found in the previous meta-analyses).

First, participants classified as aware according to the Olson and Fazio (2001) criterion showed a small evaluative-conditioning effect (Hedges's $g = 0.30$, 95% CI = [0.04, 0.56], $z = 2.23$, $p = .026$). Results from the moderator test did not provide evidence that evaluative-conditioning effects differed between aware and unaware participants, $Q(1) = 1.59$, $p = .207$. Second, participants classified as aware according to the modified Olson and Fazio (2001) criterion showed a small evaluative-conditioning effect (Hedges's $g = 0.33$, 95% CI = [0.20, 0.46], $z = 5.01$, $p < .001$). The moderator test demonstrated that evaluative-conditioning effects differed between aware and unaware participants, $Q(1) = 12.90$, $p < .001$. Third, participants classified as aware according to the original Bar-Anan et al. (2010) criterion showed a small evaluative-conditioning effect (Hedges's $g = 0.24$, 95% CI = [0.14, 0.35], $z = 4.60$, $p < .001$). The moderator test demonstrated that evaluative-conditioning effects differed between aware and unaware participants, $Q(1) = 8.10$, $p = .004$. Finally, participants classified

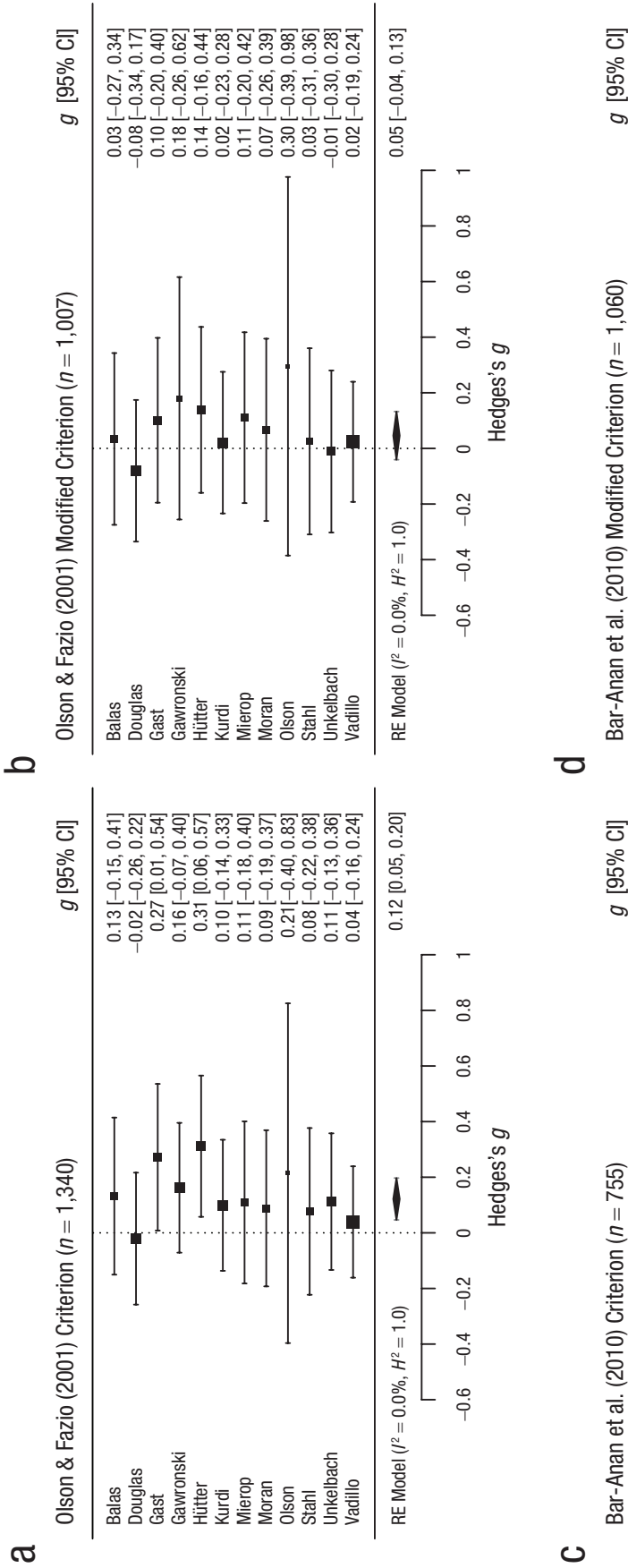


Fig. 2. Results of the preregistered meta-analytic models: effect size from each of the labs that reported results. The primary model (a) contained exclusions based on the original authors' criterion (Olson & Fazio, 2001); the secondary models contained exclusions based on the (b) Olson and Fazio (2001) modified criterion; (c) Bar-Anan, De Houwer, and Nosek (2010) criterion; and (d) Bar-Anan et al. (2010) modified criterion. The dependent variable was evaluative-conditioning effect score (i.e., a preference for positive over negative conditioned stimuli). Each lab is identified by the last name of the corresponding author. In each forest plot, squares represent observed Hedges's g effect sizes, square size represents weighting in the model (i.e., inverse variance), and error bars represent 95% confidence intervals (CIs) around the effect size. The statistics below the forest plots show the outcomes of random-effects (RE) meta-analyses. No credibility intervals beyond the CIs are visible because the CIs are visible because between-site heterogeneity was not observed. Estimates of heterogeneity (I^2 and H^2) are provided next to the meta-analysis model results. Restricted maximum likelihood estimation was used for all models.

as aware according to the modified Bar-Anan et al. (2010) criterion showed a medium evaluative-conditioning effect (Hedges's $g = 0.37$, 95% CI = [0.23, 0.51], $z = 5.24$, $p < .001$). The moderator test demonstrated that evaluative-conditioning effects differed between aware and unaware participants, $Q(1) = 14.94$, $p < .001$.

Nonpreregistered analyses: power analyses

Using the effect size found in the primary analysis and the sample sizes reported in the published literature, we found that the observed power of the Olson and Fazio (2001) study was extremely low (observed power = .13, one sample, $\alpha = .05$, two tailed), as is the observed power in the published literature on the surveillance task more generally (median power = .14, median absolute deviation = .14, range = .07–.75). This is far lower than the typically endorsed minimum power of .80 (Cohen, 1992) and out of step with the percentage of published studies that reported significant results (48%).

Using the observed effect sizes, we calculated a priori sample sizes for future research, using both the largest meta-effect size found among the four exclusion criteria (i.e., Olson & Fazio, 2001, criterion: $g = 0.12$) and the smallest (i.e., Bar-Anan et al., 2010, criterion: $g = 0.03$). Achieving 80% power would require 547 and 8,723 participants, respectively, depending on which meta-effect size is used. Achieving 95% power would require 905 and 14,441 participants, respectively. Finally, we calculated the probability of observing an effect within a sample size that is typically manageable for a single lab to collect (i.e., 150 participants: the upper bound of the recommended sample size we asked each site to collect for this article). Power analyses suggested that the probability of observing an effect (i.e., power) using a sample size of 150 was 30.9% to 6.5%, respectively, depending on which meta-effect-size estimate was used.

Discussion

Over the past 20 years, effects on the surveillance task have been treated as evidence for attitude formation in the absence of awareness/recollective memory. This claim has informed theories about evaluative conditioning and attitudes as well as interventions that are assumed to implicitly modify problematic beliefs and behavior. Yet strong claims regarding unaware evaluative conditioning require strong evidence. In this replication attempt, our primary analysis examined whether the surveillance task produced a significant evaluative-conditioning effect when the Olson and Fazio (2001) awareness exclusion criterion was used. We also conducted (preregistered) secondary analyses to investigate

whether the effect was robust under three other criteria.

Our primary analysis using Olson and Fazio's (2001) original exclusion criterion demonstrated a small but significant evaluative-conditioning effect on the surveillance task. We therefore replicated the effect, in the sense that significant results were found in both studies. However, no evaluative-conditioning effect emerged when any of the other three alternative awareness exclusion criteria were applied. To complicate matters further, evaluative-conditioning effects did not differ significantly between these four criteria. This poses a challenge in how to globally interpret effects that fall on either side of the significant versus nonsignificant divide and yet cannot be distinguished from one another in the moderator meta-analysis. Although it is correct to say that a significant evaluative-conditioning effect was found for only the primary Olson and Fazio (2001) criterion and not the other three secondary criteria, we also cannot conclude that evaluative-conditioning effects in the surveillance task depend on or differ between the specific way in which contingency awareness/recollective memory was measured, given that the difference between significant and nonsignificant effects is not itself necessarily significant. This combination of results was not covered by our preregistered plans for interpretation of results (for a detailed discussion, see the Supplemental Material).

Interpretation of the results

The failure to find significant effects with the three secondary criteria and the nonsignificant effect of exclusion-criteria type in the multilevel moderator meta-analysis creates considerable uncertainty regarding the robustness of any unaware evaluative-conditioning effect. Moreover, additional exploratory analyses conducted on the present data by some of the coauthors suggest that there is no good evidence for unaware evaluative-conditioning effects. For example, an analysis of our data that distinguishes between independent sets of fully aware, partially aware, and fully unaware participants found a nonsignificant evaluative-conditioning effect in fully unaware participants (Stahl & Corneille, 2020); a meta-analysis using a stricter compound awareness criterion that prioritized sensitivity to awareness found a nonsignificant and near-zero effect (Hussey & Hughes, 2020); and a Bayesian analysis of the data did not provide convincing evidence in favor of an unaware evaluative-conditioning effect under any of the exclusion criteria (Kurdi & Ferguson, 2020).⁶

Second, the success of a replication can also be defined in ways other than statistical significance, which may aid the interpretation of the results. Previous

large-scale replication efforts in psychology have noted a marked decrease in the effect sizes observed between original and replication studies (Open Science Collaboration, 2015). We observed a similar result here: Even the largest meta-analytic effect size among the four exclusion criteria ($g = 0.12$ using the Olson & Fazio, 2001, exclusion criterion) was approximately half that observed in the meta-analysis of published literature ($g = 0.20$) and less than half of that observed in the original study ($g = 0.27$). Results demonstrated that observed power in the published literature is therefore extremely low (median power = .14). Together, these two points suggest that the published literature on the surveillance task reports significant results at a rate far above what one should expect in the absence of publication bias or selective reporting.

Further reasons for caution can be found in the awareness concept itself. Debate continues to rage about what such exclusion criteria actually capture: Some researchers argue that it is awareness (Jones et al., 2009), whereas others advocate for recollective memory (Gawronski & Walther, 2012). For example, participants may be aware of pairings during the acquisition (evaluative-conditioning) phase but fail to recall this information during the retrieval (evaluative) phase. Although our primary analysis demonstrated that Olson and Fazio's (2001) surveillance-task effect was replicated, these conceptual concerns raise questions as to whether this procedure represents a useful test of the unaware-evaluative-conditioning hypothesis. Retrospective reports of awareness are imperfect in that they may misclassify participants as unaware or vice versa (but see Hussey & Hughes, 2020). Nonetheless, data based on retrospective measures, such as those used here, likely cannot settle the question of whether evaluative-conditioning effects can emerge in the absence of awareness by themselves. Alternative experimental manipulations of awareness are also possible; however, results from such studies also fail to produce consistent evidence of unaware evaluative conditioning (e.g., Corneille & Stahl, 2019).

The sample used in the current replication was designed to be similar to that used by Olson and Fazio (2001), in that they both employed undergraduate students. However, there are also noteworthy differences between the two samples. First, Olson and Fazio exclusively recruited female participants, whereas in the current replication, 72% of the sample were women and 28% were men. Second, whereas Olson and Fazio relied on North American participants from a single lab, the current replication recruited participants from multiple locations in North America (four labs) and Europe, the latter of which comprised non-English-speaking countries including Germany (four labs), Belgium (two labs), Spain (one lab), and Poland (one lab). Of course, reliance on undergraduate students poses a limitation to

the generalizability of both the original study and current replication's claims. However, the fact that we recruited both men and women from multiple countries and diverse language regions increases the generalizability of our findings relative to Olson and Fazio's original study.

To conclude, although we replicated the surveillance-task effect, we urge caution when using such an effect to make strong claims about unaware evaluative conditioning, especially when those claims are being used to justify new theory or interventions. We also encourage more careful reflection on existing theory and interventions that have already been founded on this effect (e.g., March et al., 2018; Shaw et al., 2016). Strong claims necessitate strong evidence—evidence that is currently lacking.

Response from Olson and Fazio

A brief response was solicited from Olson and Fazio (M. A. Olson, personal communication, June 5, 2020), and we include it here verbatim:

We emphasize that the effect was in the predicted direction in 11 of the 12 samples using the original exclusion criteria. The secondary criteria revealed analogous patterns in 10, 9, and 11 of 12 samples, respectively. However, such criteria can also exclude unaware individuals if they use their recently formed attitudes to guess CS-US valence (see Gawronski & Walther, 2012). Ultimately, the lack of a moderating effect of exclusion criteria can be interpreted as an unqualified replication of Olson and Fazio (2001).

In addition, the effect size produced by a single procedure is minimally relevant to broader theoretical questions about the multiple mechanisms that produce [evaluative conditioning]. Within our proposed implicit misattribution mechanism, the magnitude of [evaluative conditioning] is dependent [on] source confusability (the extent to which the evaluation evoked by the US is likely to be misattributed to the CS; Jones et al., 2010). Hence, future work should focus on fostering source confusability beyond the procedural parameters employed here.

Transparency

Action Editor: D. Stephen Lindsay

Editor: D. Stephen Lindsay

Author Contributions

T. Moran, S. Hughes, and I. Hussey are joint first authors of this article. T. Moran led the project administration, conducted the meta-analysis of published work, created the procedure protocol, was responsible for the design of the materials, wrote the manuscript, contributed to data

collection, and reviewed the code for the data processing and analyses. S. Hughes wrote, reviewed, and edited the manuscript and contributed to project administration. I. Hussey wrote the code for the materials, data processing, and analyses and contributed to project administration and to writing, reviewing, and editing the manuscript. M. A. Vadillo contributed to the meta-analysis of published work and to writing the original draft, writing the analyses, and reviewing and editing the final manuscript. M. A. Olson contributed to the creation of the procedure protocol, data collection, and review of the manuscript. F. Aust, K. Bading, R. Balas, T. Benedict, O. Corneille, S. B. Douglas, M. J. Ferguson, K. A. Fritzlen, A. Gast, B. Gawronski, T. Heycke, F. Högden, M. Hütter, B. Kurdi, A. Mierop, J. Richter, J. Sarzyńska-Wawer, C. T. Smith, C. Stahl, P. Thomasius, T. Giménez-Fernández, K. Hanusz, and C. Unkelbach organized or conducted data collection at their sites and contributed to the review of the manuscript. J. De Houwer contributed to the creation of the procedure protocol and review of the manuscript. All the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding





This research was conducted with the support of the following grants: Methusalem Grant BOF16/MET_V/002 of Ghent University to J. De Houwer; Ghent University Special Research Fund (BOF) Grant 01P05517 to I. Hussey; Comunidad de Madrid, Programa de Atracción de Talento Investigador Grants PSI2017-85159-P (AEI/FEDER, UE) and 2016-T1/SOC-1395 to M. A. Vadillo; Polish National Science Centre Grant UMO-2015/18/E/HS6/00765 to R. Balas; Fonds National de la Recherche Scientifique (FRS-FNRS) Grant T.0061.18 to O. Corneille; German Research Foundation (DFG) Emmy Noether Grant HU 1978/4-1 and Heisenberg Grant HU 1978/7-1 to M. Hütter; DFG Emmy Noether Grant GA 1520/2-1 to A. Gast; and DFG Grant STA 1269/3-2 to C. Stahl.








Open Practices

All data, materials, analyses, and code have been made publicly available via OSF and can be accessed at <https://osf.io/hs32y/>. All materials and analytic files were preregistered before data collection at <https://osf.io/3hjpf/>, and a deviation from the original sampling strategy was preregistered at <https://osf.io/uyng7>. Deviations from the preregistrations are discussed in the Supplemental Material available online. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

Tal Moran  <https://orcid.org/0000-0002-4681-0725>
 Ian Hussey  <https://orcid.org/0000-0001-8906-7559>
 Miguel A. Vadillo  <https://orcid.org/0000-0001-8421-816X>
 Olivier Corneille  <https://orcid.org/0000-0003-4005-4372>

Bertram Gawronski  <https://orcid.org/0000-0001-7938-3339>
 Tobias Heycke  <https://orcid.org/0000-0001-6358-6713>
 Benedek Kurdi  <https://orcid.org/0000-0001-5000-0584>
 Adrien Mierop  <https://orcid.org/0000-0002-9160-4066>
 Philine Thomasius  <https://orcid.org/0000-0002-4910-8683>
 Christian Unkelbach  <https://orcid.org/0000-0002-3793-6246>
 Jan De Houwer  <https://orcid.org/0000-0003-0488-5224>

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/0956797620968526>

Notes

1. The planned minimum sample size after 30% exclusions had 99% power to detect a Cohen's d of 0.13 and 80% power to detect a Cohen's d of 0.08 (within subjects, one tailed, $\alpha = .05$).
2. Olson and Fazio also recommended that we use mildly evocative stimuli in our replication attempt.
3. The same four neutral targets and fillers appeared with positive CSs and negative CSs.
4. Note that our preregistration and Stage 1 accepted manuscript originally referred to these as *confirmatory* versus *exploratory* analyses rather than *primary* versus *secondary*. However, this terminology was deemed to be at odds with the fact that both were preregistered and, therefore, potentially confusing for the reader. This and all other divergences from our preregistration are documented in the Supplemental Material.
5. Results from a moderator meta-analysis model that accounted for the dependency between the different exclusion criteria are reported in the Supplemental Material. This model produced similar results.
6. All commentaries related to this project are collected at <https://osf.io/qtcsw>.

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*, 2313–2335.
- Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 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*, *43*, 205–215.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.
- Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*, 161–189. doi:10.1177/1088868318763261
- De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*, Article 28046. doi: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*, 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*, 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*, 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*, 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*, 178–188.
- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*, 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*, 1161–1163.
- Hussey, I., & Hughes, S. (2020). Evaluative conditioning without awareness: Replicable effects do not equate replicable inferences. *PsyArXiv*. doi:10.31234/osf.io/4gzsp
- Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*, 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*, 933–948.
- Jones, C. R., Olson, M. A., & Fazio, R. H. (2010). Evaluative conditioning: The “how” question. In M. P. Zanna & J. M. Olson (Eds.), *Advances in Experimental Social Psychology* (Vol. 43, pp. 205–255). San Diego, CA: Academic Press.
- 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*, 1316–1321.
- Kurdi, B., & Ferguson, M. (2020). Does the surveillance paradigm provide evidence for unconscious evaluative conditioning? A Bayesian perspective. *PsyArXiv*. doi:10.31234/osf.io/n6w7c
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785.
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1995). *International Affective Picture System: Technical manual and affective ratings*. Gainesville: University of Florida.
- March, D. S., Olson, M. A., & Fazio, R. H. (2018). The implicit misattribution model of evaluative conditioning. *Social Psychological Bulletin*, *13*, Article e27574. doi:10.5964/spb.v13i3.27574
- Merckelbach, H., de Jong, P. J., Arntz, A., & Schouten, E. (1993). The role of evaluative learning and disgust sensitivity in the etiology and treatment of spider phobia. *Advances in Behaviour Research and Therapy*, *15*, 243–255.
- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*, 413–417.
- Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421–433.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. doi:10.1126/science.aac4716
- Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 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). San Diego, CA: Academic Press.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from <http://www.R-project.org>
- Schienze, A., Stark, R., & Vaitl, D. (2001). Evaluative conditioning: A possible explanation for the acquisition of disgust responses? *Learning and Motivation*, *32*, 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., & Corneille, O. (2020). Evaluative conditioning in the surveillance paradigm is moderated by awareness exclusion criteria. *PsyArXiv*. doi:10.31234/osf.io/3xsbu
- 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*, 404–420.
- Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., . . . Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, Article d4002. doi:10.1136/bmj.d4002
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). doi:10.18637/jss.v036.i03
- Walther, E., Nagengast, B., & Trasselli, C. (2005). Evaluative conditioning in social psychology: Facts and speculations. *Cognition and Emotion*, *19*, 175–196.