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

Tal Moran[*], Sean Hughes[*], Ian Hussey[*], Miguel A. Vadillo, Michael A. Olson, Frederik Aust, Karoline Bading, Robert Balas, Taylor Benedict, Olivier Corneille, Samantha B. Douglas, Melissa J. Ferguson, Katherine A. Fritzlen, Anne Gast, Bertram Gawronski, Tamara Giménez-Fernández, Krzysztof Hanusz, Tobias Heycke, Fabia Högden, Mandy Hütter, Benedek Kurdi, Adrien Mierop, Jasmin Richter, Justyna Sarzyńska-Wawer, Colin Tucker Smith, Christoph Stahl, Philine Thomasius, Christian Unkelbach & Jan De Houwer

Evaluative conditioning (EC) is one of the most widely-studied procedures for establishing and changing attitudes. The surveillance-task (Olson & Fazio, 2001) is a highly cited EC paradigm, and one that is claimed to generate attitudes without awareness. The potential for EC 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 ($N = 1478$ adult participants), preregistered close replication of the original surveillance-task study. We obtained evidence for a small EC 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' EC effects.

Evaluative conditioning (EC) is a widely-studied and highly applicable procedure for establishing and changing attitudes (e.g., De Houwer et al., 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 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 (APE) Model (Gawronski & Bodenhausen, 2006), an influential theory of attitudes in social psychology, distinguishes between explicit attitudes and implicit attitudes, and treats 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 et al., 2001), phobias (e.g., Merckelbach et al., 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 et al., 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, debate is largely led by proponents of dual-process (e.g., Gawronski & Bodenhausen, 2006), single process

[*] Joint first authors

propositional (e.g., De Houwer, 2018), and associative 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, 2019). Showing that EC effects can occur without contingency awareness is often viewed as supporting 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 EC effects are highly dependent on contingency awareness (e.g., Bar-Anan et al., 2010; Hofmann et al., 2010; Stahl et al., 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).

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 conditioned stimuli. Valenced pictures and words serve as unconditioned stimuli. Participants are told that they will take part in a 'surveillance task' wherein they have to detect several target Pokémon that are different to 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 (US positive) whereas on other 'distractor' trials a second Pokémon (CS2) is always presented with a negative word or image (US negative). In this way, the task requires people to process the CS-US pairs but directs their attention away from those pairings and towards the irrelevant target stimuli. Afterwards, relative preferences for CS1 and CS2 is 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, 2010; March et al., 2018).

Since its introduction in 2001, the surveillance task has become one of the most frequently cited EC procedures in the literature (over 700 citations in Google Scholar as of June 2020). Several authors have claimed that the surveillance task provides evidence 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 et al., 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).

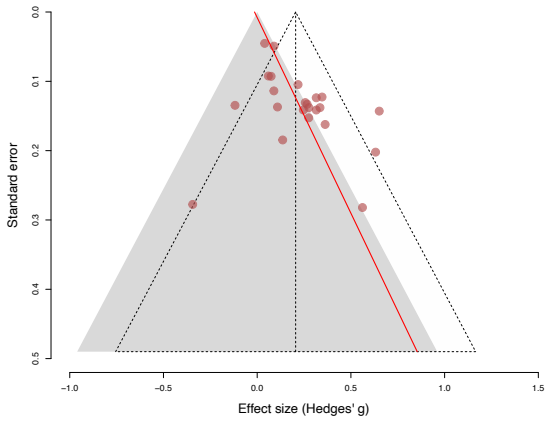

Figure 1. Funnel plot of the data entered into the meta-analysis of previous studies with the surveillance task. Each dot depicts an effect size (Hedges' g) against their Standard Errors. Studies falling inside the grey area are statistically non-significant in a two-tailed test. The triangle inside 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 bold line represents Egger's regression test for funnel plot asymmetry.

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 supported the possibility of EC effects without awareness/recollective memory using the surveillance paradigm. A random-effects meta-analysis of these studies (see osf.io/4mh2d) 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 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, $\chi^2(1) = 6.49$, $p = .011$, 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 argued to provide evidence for EC effects without awareness/recollective memory, is used to advocate for dual-process and associative 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 task's use 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 directly replicating their original design, they encouraged us to make 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). It is therefore important to note that this study represents a close conceptual replication rather than a direct replication of Olson and Fazio (2001). The final study protocol was approved by the original authors (see osf.io/wnckg). 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 criterion may have inadvertently included individuals who were aware of/remembering the contingencies. We therefore included three additional contingency awareness/recollective memory measures that assess this construct in a more conservative manner.

Disclosures

All materials and analytic files were preregistered before data collection (see osf.io/3hjpf and osf.io/uyng7). All materials, data, analyses, and code are available on the Open Science Framework (osf.io/hs32y). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data was 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

1478 adult participants were recruited from twelve labs at ten universities in Europe and North America (72% women, 27% men, < 1% other identity, $M_{age} = 21.2$ $SD = 4.9$). All labs used an ad hoc sampling strategy to sample from undergraduate students, and all experimental sessions were run in person (i.e., rather than online). We initially planned that each lab would collect data from a minimum of 100 participants and a maximum of 150 participants based on their local resources. The rationale for this planned sample size was that in previously published studies the proportion of contingency aware participants ranged from 2% to 27%. Consequently, 1200 participants would allow for greater than 99% power to observe a small EC effect (Cohen's $d = 0.20$) even if 30% of the sample were subsequently excluded on the basis of contingency awareness/recollective memory.[2] See the Supplemental Online Material-Reviewed (SOM-R: osf.io/z2vts) for details on the sample size and characteristics for each lab. All data from all sites was included in the analyses, following the amended preregistration for our data collection stopping rule (osf.io/uyng7).

Materials

Unconditioned stimuli. Study materials provided by the original authors were used. Ten positive words, ten negative words, ten positive images, and ten negative images served 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*) were identical to those used in Experiment 5 of Jones et al. (2009).[3] The

[2] The planned minimum sample size after 30% exclusions had 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).

[3] The original authors also recommended that we use mildly evocative stimuli in our replication attempt.

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 the CSs from their original (2001) 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. Based on 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) which differed least in valence and familiarity served as CSs (see the SOM-R for more details osf.io/z2vtsand osf.io/a3qj9 for the results of the pretest conducted at each lab).

Filler and target stimuli. The seven characters not selected during the pre-testing 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. The original authors did not provide us with filler items and we had to therefore select these items and have them approved by those authors.

Procedure

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

Surveillance task. The surveillance task consisted of 5 blocks, each containing a different target stimulus. Each block comprised of 86 trials, each presented for 1500ms with no inter-trial interval. Each block included 8 CS-US pair trials (4 CS-US$_{pos}$ trials and 4 CS-US$_{neg}$ 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, 80-82, with an alternation between the CS$_{pos}$ and CS$_{neg}$). 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 (approximately 1cm 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 (see osf.io/wnckg for a detailed overview of trial content). Prior to the surveillance task participants were instructed to detect the target stimulus and hit the space-bar every time a target stimulus appeared (see osf.io/wnckg for the specific instructions).

Filler task. Although a filler task was not used in the original (2001) study nor in the vast majority of published surveillance task studies (4 of the 23 studies in our meta-analysis), the original authors recommended that we add a filler task in order to create a delay between the surveillance task and the evaluation task (e.g., Kendrick & Olson, 2012). The filler task included two questionnaires: the Need for Cognition scale (18-item NFC Scale: Cacioppo et al., 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 and were therefore not analyzed. Nevertheless, those interested in these data can retrieve it from the OSF website (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 onscreen 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 CS$_{pos}$ and CS$_{neg}$ together, four presented the CS$_{pos}$ with one of the neutral targets/fillers, and four presented CS$_{neg}$ with one of the neutral targets/fillers). [4] The remaining 20 trials were filler trials, each presenting two neutral targets/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 ten critical trials were randomly assigned to the fixed positions (see osf.io/wnckg for the instructions preceding the evaluation task).

Post-experiment questionnaire. After the evaluation task, participants completed a questionnaire: we used the original Olson and Fazio (2001) post-experiment questionnaire followed by the questionnaire used in the studies of Bar-Anan et al. (2010). With respect to the former, participants first answered three open-ended questions: 1. *Think back to the very first part of the experiment. Did you notice anything out of*

[4] The same four neutral targets/fillers appeared with CS$_{pos}$ and CS$_{neg}$.

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 were 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: CS_{pos} (certainly), CS_{pos} (probably), CS_{pos} (guess), CS_{neg} (guess), CS_{neg} (probably), CS_{neg} (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 in order 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 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 non-target trials, or not responding to target trials), to check if participants paid attention during that task. Based on the original authors' recommendations, we excluded participants who were more than three standard deviations above or below the mean number of errors, as in the original Olson and Fazio (2001) study. 2% of participants were excluded on this basis.

Evaluation task. Following Jones et al. (2009), a *self-reported preference score* was calculated for each participant based on their performance during the evaluation task. Specifically, a score of 1 was assigned to trials in which the participant chose the CS_{pos} or the image that appeared together with CS_{neg}. A score of −1 was assigned to trials in which participants chose the CS_{neg} or the image appearing together with CS_{pos}. The sum of this coding, which ranged from -10 to +10 served as a measure of evaluative responding (i.e., a preference for CS_{pos} over CS_{neg}).

Awareness/recollection memory criteria. Four methods of excluding individuals based on their responses to the post-experimental questions were preregistered. The first was similar to that employed by the original authors in their study (Olson & Fazio, 2001), whereas the other three were included to explore the robustness of the effect. These latter criteria had either been used in previously published work (Bar-Anan et al., 2010), or were created by us to provide different levels of stringency around awareness than previously employed (i.e., higher than Olson & Fazio, 2001 and lower than the Bar-Anan et al., 2010).[5]

Primary criterion: Olson and Fazio (2001). A score was computed following the original authors' recommendations. This score was based on participants' open-ended responses to the original Olson and Fazio (2001) post-experimental questions 1 and 2 (see SOM-R for more detailsosf.io/z2vts). Two independent raters, who were blinded to one another's ratings, evaluated responses to these two questions, and treated responses on both questions as one (compound) text response (see osf.io/2dm6u for the exact coding instructions provided to the data collection sites). 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 only scored as 'aware' if both raters scored them as 'aware'.

[5] Note that our preregistration and Stage 1 accepted manuscript originally referred to these as 'confirmatory' vs. 'exploratory' analyses rather than 'primary' vs. '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 preregistration are documented in the SOM-R.

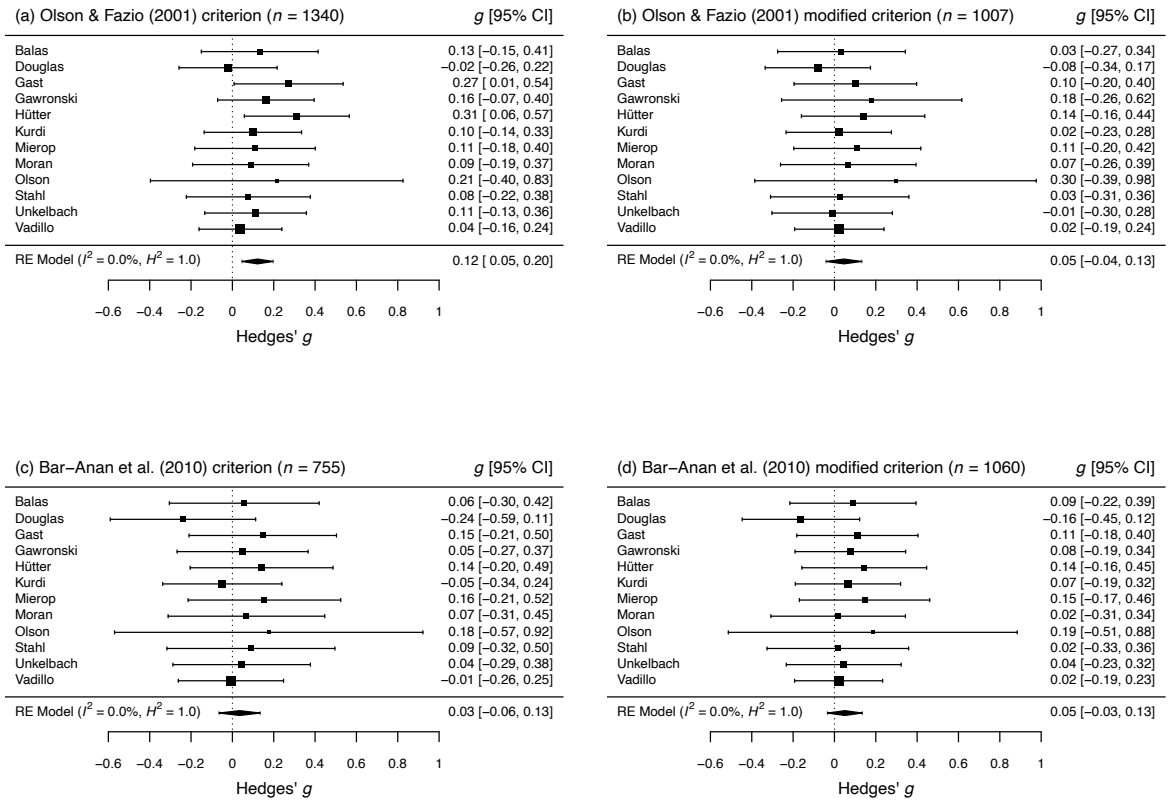

Figure 2. Results of the preregistered meta-analytic models. Primary model: (a) exclusions based on the original authors' criterion (Olson & Fazio, 2001); secondary models: exclusions based on the (b) Olson & Fazio (2001) modified, (c) Bar-Anan et al. (2010) and (d) Bar-Anan et al. (2010) modified criteria. DV was Evaluative Conditioning effect score (i.e., a preference for CS_{pos} over CS_{neg}). Each lab is identified by the last name of the corresponding author. In each forest plot, squares represent observed Hedges' *g* effect sizes, size of square represents weighting in the model (i.e., inverse variance), and error bars represent 95% Confidence Intervals (CIs) around the effect size. The bottom row in the figure is the outcome of a random-effects meta-analysis. No credibility intervals beyond the confidence intervals are visible due to no between site heterogeneity being 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.

Secondary criteria. The original authors' 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 if evidence for EC effects in this task were robust to, or depended on, the specific way in which contingency awareness/recollective memory was measured. As detailed in SOM-R (osf.io/z2vts), 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 to 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 EC 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 EC effect size (Hedges' *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 value of 0.05 (two-sided). Although all labs used similar materials, they may nevertheless differ in the translation of materials, selection of stimuli, or characteristics of the samples. In order 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.

EC effects in the absence of contingency awareness/recollective memory.

Primary analyses. The meta-analysis based on the original Olson and Fazio (2001) awareness criterion

($n = 1340$, 9.2% excluded) showed that, on average, the surveillance task led to a small but significant EC effect, Hedges' $g = 0.12$, 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 Figure 2, panel 'a'). Variation in effect sizes between sites was consistent with what one would expect by chance (i.e., due to sampling variation alone), $\tau^2 = 0.0$, $I^2 = 0.0\%$, $H^2 = 1.0$, $Q(11) = 5.83$, $p = .885$. In sum, when the original authors' awareness exclusion criterion was employed, their original effect was replicated.

Secondary analyses. When a modified version of the original authors' exclusion criterion was applied (i.e., Olson & Fazio, 2001 modified, $n = 1007$, 31.9% excluded), the surveillance task was not found to produce an EC effect, Hedges' $g = 0.05$, 95% CI [-0.04, 0.13], $z = 1.04$, $p = .299$. Effect sizes ranged from -0.08 to 0.30 between sites (see Figure 2, panel 'b'). Variation in effect sizes between 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 EC effect, Hedges' $g = 0.03$, 95% CI [-0.06, 0.13], $z = 0.69$, $p = .493$. Effect sizes ranged from -0.24 to 0.18 between sites (see Figure 2, panel 'c'). Variation in effect sizes between 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 ($n = 1060$, 28.3% excluded), the surveillance task did not lead to an EC effect, Hedges' $g = 0.05$, 95% CI [-0.03, 0.13], $z = 1.17$, $p = .241$. Effect sizes ranged from -0.16 to 0.19 between sites (see Figure 2, panel 'd'). Variation in effect sizes between 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 based on the four awareness/recollective memory criteria differ from one another, we combined the datasets 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 between effect sizes coming from related samples. The moderator test did not demonstrate evidence that the results of the four criteria differed from each other, $Q(3) = 2.76$, $p = .430$.[6]

Comparison of 'contingency-aware' vs. 'unaware' participants. The previous analyses excluded 'contingency-aware' participants. Yet one could also examine whether awareness/recollective memory moderate the size of EC 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 EC effects between these two groups using a multilevel moderator meta-analysis model (see SOM-R for more details about these analyses osf.io/z2vts). All moderator analyses reported in this section included a random intercept for data collection site in order 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 EC effect, Hedges' $g = 0.30$, 95% CI [0.04, 0.56], $z = 2.23$, $p = .026$. Results from the moderator test did not provide evidence that EC 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 EC effect, Hedges' $g = 0.33$, 95% CI [0.20, 0.46], $z = 5.01$, $p < .001$. The moderator test demonstrated that EC 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 EC effect, Hedges' $g = 0.24$, 95% CI [0.14, 0.35], $z = 4.60$, $p < .001$. The moderator test demonstrated that EC effects differed between 'aware' and 'unaware' participants, $Q(1) = 8.10$, $p = .004$. Finally, participants classified as 'aware' according to the modified Bar-Anan et al. (2010) criterion showed a medium EC effect, Hedges' $g = 0.37$, 95% CI [0.23, 0.51], $z = 5.24$, $p < .001$. The moderator test demonstrated that EC effects differed between 'aware' and 'unaware' participants, $Q(1) = 14.94$, $p < .001$.

Non-Preregistered Analyses: Power Analyses

Using the effect size found in the primary analysis and the sample sizes reported in the published literature, the observed power of the original Olson and Fazio (2001) study was extremely low (observed power = .13, one-sample, alpha = 0.05, two-sided), as is the observed power for the published literature on the surveillance task more generally (median power = .14, MAD = .14, range = .07 to .75). This is far lower than the typically endorsed minimum of power = .80 (Cohen, 1992), and out of step with the proportion of published studies that reported significant results (48%).

[6] Results from a moderator meta-analysis model that accounts for the dependency between the different exclusion criteria are reported in SOM-R (osf.io/z2vts). This model produced similar results.

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: $g = 0.03$). To achieve 80% power, $n > 547$ to 8723 participants would be required, respectively, depending on which meta-effect size is used. To achieve 95% power, $n > 905$ to 14,441 participants would be required, 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 the probability of observing an effect (i.e., power) using a sample size of $n = 150$ was 30.9% to 6.5% respectively, depending on which meta-effect size estimate was used.

Discussion

Over the past twenty 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 EC and attitudes, as well as interventions that are assumed to 'implicitly' modify problematic beliefs and behavior. Yet strong claims regarding 'unaware' EC require strong evidence. In this replication attempt, our *primary* analysis examined whether the surveillance task produced a significant EC effect when the original 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 EC effect on the surveillance task. We therefore replicated the effect, in the sense that significant results were found in both studies. However, no EC effect emerged when any of the other three alternative awareness exclusion criteria were applied. To complicate matters further, EC effects did not differ significantly between these four criteria. This poses a challenge in how to make a global interpretation of effects that (a) fall on either side of the significant versus non-significant divide, and yet (b) cannot be distinguished from one another in the moderator meta-analysis. While it is correct to say that a significant EC effect was found for only the primary Olson and Fazio (2001) criterion and not the other three secondary criteria, we also cannot conclude that EC 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 non-significant effects is not itself necessarily significant. This combination of results was not covered by our preregistered plans for interpretation of results (for detailed discussion see SOM-R: osf.io/z2vts).

Interpretation of the Results

The failure to find significant effects with the three secondary criteria and the non-significant effect of exclusion criteria type in the multilevel moderator meta-analysis creates considerable uncertainty regarding the robustness of any 'unaware' EC effect. Moreover, additional exploratory analyses conducted on the present data by some of the co-authors suggest that there is no good evidence for 'unaware' EC effects. For example, an analysis of our data that distinguishes between independent sets of 'fully aware', 'partially aware', and 'fully unaware' participants found a non-significant EC effect in 'fully unaware' participants (Stahl & Corneille, 2020); a meta-analysis using a stricter compound awareness criterion that prioritized sensitivity to awareness found a non-significant and near-zero effect (Hussey & Hughes, 2020); and a Bayesian analysis of the data did not provide convincing evidence in favor of 'unaware' EC effect under any of the exclusion criteria (Kurdi & Ferguson, 2020).[7]

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 that we observed 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 = 0.14). 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 argue that it is 'awareness' (Jones et al., 2009) whereas others advocate for 'recollective memory' (Gawronski & Walther, 2012). For example,

[7] All commentaries related to this project are collected at osf.io/qtcsw.

participants may be aware of pairings during the acquisition (EC) 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' EC 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 EC 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' EC (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 were comprised of non-English speaking countries including Germany (four labs), Belgium (two labs), Spain (1 lab) and Poland (1 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 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' EC, 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 the Original Authors

A brief response was solicited from the original authors and we include it here verbatim. "We [Olson and Fazio] 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 EC. Within our proposed implicit misattribution mechanism, the magnitude of EC is dependent upon 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."

Author contributions

TM led the project administration, conducted the meta-analysis of published work, created the procedure protocol, was responsible for design of the materials, wrote the manuscript, contributed to data collection, and reviewed the code for the data processing and analyses. SH wrote, reviewed, and edited the manuscript and contributed to project administration. IH wrote the code for the materials, data processing, and analyses, and contributed to project administration, and writing, reviewing, and editing the manuscript. MAV contributed to the meta-analysis of published work, and to writing the original draft, the analyses, and reviewing and editing the final manuscript. MAO contributed to the creation of the procedure protocol, data collection and review of the manuscript. FA, KB, RB, TB, OC, SBD, MJF, KAF, AG, BG, TH, FH, MH, BK, AM, JR, JSW, CTS, CS, PT, TGF, KH and CU organized and/or conducted data collection at their sites, and contributed to the review of the manuscript. JDH contributed to the creation of the procedure protocol and review of the manuscript.

Funding

This research was conducted with the support of the following grants: FWO grant BOF16/MET_V/002 to Jan De Houwer, Ghent University BOF grant 01P05517 to Ian Hussey, Comunidad de Madrid, Programa de Atracción de Talento Investigador grants PSI2017-85159-P (AEI / FEDER, UE) and 2016-T1/SOC-1395 to Miguel Vadillo, Polish National Science Centre grant UMO-2015/18/E/HS6/00765 to Robert Balas, FRS-FNRS grant T.0061.18 to Olivier Corneille, DFG Emmy Noether grant HU 1978/4-1 and Heisenberg grant HU 1978/7-1 to Mandy Hütter, DFG-Emmy-Noether-Grant GA 1520/2-1 to Anne Gast, and DFG grant STA 1269/3-2 to Christoph Stahl.

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.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 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*(2), 161-189. doi.org/10.1177/1088868318763261

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13*(3), e28046. 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*(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.

Hussey, I., & Hughes, S. (2020). Evaluative Conditioning without awareness: Replicable effects do not equate replicable inferences. Preprint. https://psyarxiv.com/4gzsp/

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.

Kurdi, B., & Ferguson, M. (2020). Does the surveillance paradigm provide evidence for unconscious evaluative conditioning? A Bayesian perspective. Preprint. https://psyarxiv.com/n6w7c/

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. (1997). International Affective 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.

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*(4), 243–255.

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.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

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., & Corneille, O. (2020). Evaluative conditioning in the Surveillance paradigm is moderated by awareness exclusion criteria. Preprint. https://psyarxiv.com/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*(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.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48. 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*(2), 175-196.