INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Unaware Attitude Formation in the Surveillance Task? Revisiting the Findings of Moran et al. (2021)

RESEARCH ARTICLE

BENEDEK KURDI
IAN HUSSEY
CHRISTOPH STAHL
SEAN HUGHES
CHRISTIAN UNKELBACH
MELISSA J. FERGUSON
OLIVIER CORNEILLE

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Moran et al. (2021) report a multi-lab registered replication of Olson and Fazio's (2001) surveillance task. The surveillance task is an incidental learning procedure over the course of which participants observe pairings of conditioned stimuli (CSs) and unconditioned stimuli (USs) while engaging in a distracting secondary task. Unaware evaluative conditioning (EC) effects are inferred if participants who fail to report the CS–US contingencies on a post-hoc measure show preference for the CS_{pos} over the CS_{neg}. Moran et al. claimed to establish such effects relying on the criteria used by Olson and Fazio to exclude contingency aware participants from analyses. Here we reexamine Moran et al.'s data using more fine-grained analytic strategies. We show that the contingency awareness measures used by Olson and Fazio and, by extension, Moran et al. lack adequate reliability and validity. Moreover, even assuming valid awareness measures, Bayesian analyses did not provide unambiguous evidence for unaware EC effects under any exclusion criterion and provided decisive evidence against such effects in most models. Finally, a separate analysis that distinguished between fully aware, partially aware, and fully unaware participants shows that evidence for unaware EC is due to the inclusion of partially aware participants in the purportedly unaware subsample. These reanalyses suggest that unaware EC as indexed by the surveillance task has yet to be convincingly demonstrated. We discuss the conceptual, theoretical, and applied implications of these findings with regard to the potential for unaware attitude formation.

CORRESPONDING AUTHOR:
Benedek Kurdi
Yale University, US
benedek.kurdi@yale.edu

KEYWORDS:
attitudes; awareness; evaluative conditioning; incidental learning; surveillance task

TO CITE THIS ARTICLE:
Kurdi, B., Hussey, I., Stahl, C., Hughes, S., Unkelbach, C., Ferguson, M. J., & Corneille, O. (2022). Unaware Attitude Formation in the Surveillance Task? Revisiting the Findings of Moran et al. (2021). *International Review of Social Psychology,* 35(1): 6, 1–16. DOI: https://doi.org/10.5334/irsp.546

The study of attitudes, or the tendency to evaluate an entity with a certain degree of favor or disfavor (Eagly & Chaiken, 1993), has been a topic of inquiry in scientific psychology since the very inception of the field. Within the broad area of attitude research, interest in the origins and properties of attitudes has remained remarkably constant over the past century. One focal question has centered on attitude acquisition and change. Here, we address one particular facet of this issue: Can attitudes form and change in the absence of awareness?

The topic of unaware attitude acquisition and change has long intrigued attitude researchers. Dual-process accounts suggest that such effects are possible and argue that such learning is most likely to occur when evaluations are established or changed in purportedly simple ways, such as via conditioning or mere exposure (e.g., Gawronski & Bodenhausen, 2014; Rydell et al., 2006). The idea that evaluations can be formed and revised without awareness is also central to implicit misattribution models (e.g., Jones et al., 2009) and in the domain of implicit social cognition (e.g., Greenwald & Banaji, 1995).

In contrast, more recent propositional perspectives on learning (e.g., Mitchell et al., 2009), and evaluative learning in particular (e.g., De Houwer, 2014), emphasize the role of certain conditions, including, notably, awareness, in the formation of propositional representations about stimulus relations, even in seemingly simple paradigms such as mere exposure or evaluative conditioning. As such, whether attitude change can occur in the absence of awareness has been a much investigated and controversial topic in evaluative learning research over the past decades.

One paradigm that has often been used to study the role of awareness in attitude acquisition and change is evaluative conditioning (EC). In EC procedures, a neutral conditioned stimulus (CS) acquires the valence of a positive or negative unconditioned stimulus (US) following exposure to pairings of the two. Given the apparent simplicity of this procedure, EC effects have long been assumed to be mediated by 'simple' (associative) mental mechanisms, and to potentially occur in the absence of awareness (e.g., Baeyens et al., 2009; Jones et al., 2009; Levey & Martin, 1975). In this context, 'absence of awareness' generally refers to the absence of conscious encoding of the CS–US pairings, that is, an absence of contingency awareness. In correlational studies, the absence of memory for the CS–US pairings is considered a proxy for such lack of awareness (a point to which we return below).

However, compelling evidence for unaware EC is currently lacking. A meta-analysis by Hofmann et al. (2010) supports the role of recollective memory of CS–US contingencies in EC effects. Likewise, experimental work that directly prevents conscious encoding of the CS–US pairings during learning has largely failed to obtain evidence for EC effects. This was, for instance, the case in experiments involving brief (Stahl et al., 2016), visually suppressed (Högden et al., 2018), and parafoveal (Dedonder et al., 2014) stimulus presentations. Likewise, an EC effect does not emerge when participants' cognitive resources are depleted during learning—a manipulation that disrupts conscious encoding of the CS–US pairings (e.g., Davies et al., 2018; Dedonder et al., 2010; Kattner, 2012; for a review, see Corneille & Stahl, 2019).

THE SURVEILLANCE TASK AND THE REPLICATION BY MORAN ET AL. (2021)

To summarize, robust experimental evidence for unaware EC effects is not currently available. Yet, one particular paradigm and the results obtained using this paradigm are frequently cited in support of the idea of unaware EC: the *surveillance task* introduced by Olson and Fazio (2001). In this paradigm, participants are asked to assume the role of security guard and to monitor the presence of specific images and words in a stream of stimuli appearing on the computer screen across multiple blocks. Unbeknownst to participants, trials seemingly irrelevant to their primary task of tracking the appearance of certain stimuli include systematic pairings of two Pokémon characters with valenced words and images.

Specifically, one initially neutral Pokémon stimulus (CS_{pos}) is consistently paired with positive words and images (US_{pos}) and a second neutral Pokémon stimulus (CS_{neg}) with negative words and images (US_{neg}). Following such training, participants have been found to show a preference for the CS_{pos} over the CS_{neg} (i.e., an EC effect), including participants who fail to report the CS–US contingencies when retrospectively asked about them after the study. As mentioned above, the fact that changes in liking occur despite a lack of retrospective self-report of the CS–US pairings has been interpreted as evidence that EC effects can emerge in the absence of awareness.

Critically, even setting aside the more general issue of whether lack of awareness can be inferred from lack of retrospective memory (see below), the validity of the conclusions regarding unaware EC in the surveillance task hinges on the specificity and sensitivity of the measures used to exclude participants deemed as contingency aware (Lovibond & Shanks, 2002; Newell & Shanks, 2014; Shanks & St. John, 1994). That is, the measure used to establish contingency awareness should (*a*) identify all participants that were contingency aware as contingency aware and (*b*) not erroneously identify any participants that were contingency unaware as contingency aware. In the context of the surveillance task, false negatives are particularly problematic: If the awareness measure were to misclassify large numbers of

contingency aware participants as contingency unaware, such misclassification could give rise to erroneous claims of unaware attitude formation.

The criterion used by Olson and Fazio (2001) to exclude contingency aware participants is worth critically reexamining in this regard. Specifically, at the end of the experiment, participants in the Olson and Fazio (2001) studies were asked to respond to two open-ended questions ('Did you notice anything out of the ordinary in the way the words and pictures were presented during the surveillance tasks?' and 'Did you notice anything systematic about how particular words and images appeared together during the surveillance tasks?'). Participants were excluded as contingency aware only if they correctly reported both CS–US pairings, that is, pairings of the CS_{pos} with the US_{pos} and pairings of the CS_{neg} with the US_{neg}, in response to these items. All remaining participants were scored as unaware, including participants who (*a*) identified only one of the two CS–US pairings, (*b*) reversed the CS_{pos} with the CS_{neg} in their answer, or (*c*) mentioned CS–US pairings but did not mention the valence of the USs. Arguably, given the expectation of a highly specific description in response to open-ended questions, the use of these items is associated with a considerable risk of misclassifying participants who are (partially) aware of the CS–US contingencies as contingency unaware.

In a recent registered replication report (RRR) relying on data from a large sample of participants (N = 1,478) collected across 12 laboratories in nine European countries and the United States, Moran et al. (2021) set out to investigate the replicability of and potential boundary conditions on the surveillance task effect. Given concerns about the sensitivity of the Olson and Fazio (2001) exclusion criterion, Moran and colleagues considered three secondary sets of exclusion criteria. Specifically, the modified Olson and Fazio (2001) criterion considered participants to be contingency aware if they mentioned any systematic pairings between CSs and USs in response to the items cited above, including identification of only one of two sets of CS–US pairings and the mention of systematic pairings without identifying the valence of the USs.

In addition to these criteria, Moran et al. (2021) also adapted contingency awareness items from Bar-Anan et al. (2010). Unlike the Olson and Fazio (2001) criteria, the Bar-Anan et al. (2010) items did not require hand coding of responses. Specifically, the original Bar-Anan et al. (2010) criterion asked, '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?' A modified version of the same item asked participants to correctly identify the CS_{pos} and the CS_{neg} from the specific set of CSs to which they had been exposed. Participants who selected the correct response on both items and indicated a confidence level above guessing ('probably' or 'certainly') were classified as contingency aware.

The Moran et al. (2021) RRR yielded mixed results, including inconsistencies as a function of the criteria selected to identify and exclude contingency aware participants. Specifically, when Moran and colleagues meta-analyzed all previously published surveillance task effects, they found a small but statistically significant effect when publication bias was not corrected for. When publication bias was corrected for, the effect disappeared. Using the data obtained in the multi-lab replication study itself, a small but statistically significant EC effect was detected when Olson and Fazio's (2001) original criterion was used to exclude contingency aware participants prior to analyses. No such effect emerged when three alternative exclusion criteria were used: the modified version of the Olson and Fazio (2001) criterion, and original and modified exclusion criteria based on Bar-Anan et al. (2010).

In addition, to complicate matters even further, although statistical significance of the unaware EC effect differed across criteria, the difference across criteria itself did not reach statistical significance in a moderator analysis conducted by Moran et al. (2021). This set of findings allows for multiple conflicting interpretations of the data. Whereas Olson and Fazio viewed these outcomes as supporting their perspective (interpreting the results as 'unqualified' evidence for a successful replication), some co-authors viewed the same findings as providing evidence against the idea that the surveillance task produces unaware EC effects. Here, we revisit these conflicting interpretations.

THE PRESENT WORK: A THREE-TIERED APPROACH TO THE VALIDITY OF THE SURVEILLANCE TASK

In the present work we take a three-tiered approach toward examining the validity of inferences about unaware EC effects in the surveillance task. The first tier concerns the validity of correlational approaches to contingency awareness in general; the second tier concerns the validity of the specific contingency awareness measures used by Olson and Fazio (2001) and, by extension, by Moran et al. (2021), assuming that the correlational approach is valid; and the third tier concerns the validity of the statistical inferences regarding the presence of unaware EC effects in the surveillance task, assuming that the particular measures of awareness used by Olson and Fazio (2001) are valid. In other words, each tier of validity analysis makes increasing concessions toward accommodating the perspective of Olson and Fazio (2001; and, by extension, Moran et al., 2021). And yet, to anticipate our findings,

we did not obtain compelling evidence for unaware EC effects at any tier of analysis.

In our main analyses, we reexamined the Moran et al. data using more fine-grained analytic strategies than the original authors. In a first set of analyses, we asked whether, putting aside more general conceptual difficulties inherent to the correlational approach to unaware EC, the reliability and validity of the particular contingency awareness items used by Olson and Fazio (2001) and, by extension, by Moran et al. (2021) is appropriate. In the remaining two sets of analyses, we relaxed assumptions even further by taking the validity of the contingency awareness measure at face value.

Specifically, we used Bayesian modeling to probe whether the results emerging from the surveillance task are sensitive to (*a*) the exclusion criteria selected and (*b*) the choice of prior, including the meta-analytic estimate reported by Moran et al. (2021) and the same meta-analytic estimate adjusted for publication bias. Finally, we fit an additional set of meta-analytic models to the data that distinguish between independent sets of (*a*) 'fully aware,' (*b*) 'partially aware,' and (*c*) 'fully unaware' participants. Although both of these sections are concerned with the validity of statistical inferences, the former focuses on the robustness of the results to different specifications of the Bayesian model, whereas the latter focuses on dependencies between exclusion criteria.

VALIDITY OF THE CORRELATIONAL APPROACH TO CONTINGENCY AWARENESS

Before turning to our main analyses, we address the general issue of using post-hoc measures of retrospective memory, as implemented in the surveillance task, to establish learning in the absence of contingency awareness. Given that this is a conceptual question with substantial relevant theorizing in previous work (e.g., Corneille & Stahl, 2019; Gawronski & Walther, 2012; Sweldens et al., 2014), here we limit ourselves to a brief summary. Specifically, prior theoretical work has pointed out the inherently dubious nature of inferring (lack of) awareness of CS–US contingencies from a test of retrospective memory.

Notably, participants who were aware of such contingencies during learning may not be able or willing to report them at test for multiple reasons: For example, given that considerable amounts of time can elapse between the learning phase and the end of the test phase (when contingency awareness measures are usually administered), forgetting or memory interference may impede correct responding. In addition, participants may misinterpret the contingency awareness items. This consideration is paramount when it comes to open-ended measures, such as the one used by Olson and Fazio (2001). To name just one potential point of confusion, participants may not have considered systematic CS–US pairings as being 'out of the ordinary.' Finally, given that contingency awareness measures are usually included at the end of the experiment, participants may not be motivated to respond accurately; rather, they may prefer to reach the end of the study as quickly as possible.

As such, using retrospective measures of contingency memory to provide evidence against contingency awareness seems conceptually problematic. However, we believe that the results emerging from the surveillance task and the recent multi-lab RRR should still be considered informative, for multiple reasons. Notably, the Olson and Fazio (2001) paper is a classic with over 800 citations as of this writing. Therefore, whether its findings are numerically replicable, even if theoretically ambiguous, is of inherent interest. Moreover, the findings emerging from this paradigm are directly relevant to the question of whether evaluative learning can occur in the absence of contingency memory, although this question is not identical to the question of whether it can occur in the absence of contingency awareness. Finally, even if the overall approach taken by a study could be considered questionable, we see value in accepting the original approach at face value and asking whether the conclusions of that study seem robust and replicable within the confines of that particular approach.

VALIDITY OF THE OLSON AND FAZIO (2001) CONTINGENCY AWARENESS MEASURES

Testing the unaware EC hypothesis requires a reliable and valid measure capable of excluding participants who were aware of the CS–US pairings. Here we consider the reliability and validity of the awareness exclusion criteria used by Olson and Fazio (2001; and, by extension, Moran et al., 2021). Although awareness measures are frequently unreliable (Shanks, 2017; Vadillo et al., 2019), neither the original article nor the RRR directly considered this problem. Recent work has argued that such issues related to measurement are common yet underappreciated in psychology and can threaten the validity of findings and the conclusions researchers draw from them (e.g., Flake et al., 2017; Flake & Fried, 2020; Hussey & Hughes, 2019).

At least in part, the effect obtained in Moran et al.'s (2021) primary analysis seems to have been driven by the fact that the exclusion criterion used in that analysis failed to exclude individuals who were aware of the CS–US pairings. As such, here we (*a*) assess the validity and reliability of the four awareness criteria and, using

differences in responding to different measures of awareness across individuals and across data collection sites, conclude that they are poor and noisy measures of awareness and (*b*) conduct a stricter test of the core verbal hypothesis and conclude that evidence for unaware EC is explained by the inclusion of partially aware participants in the unaware group.

ASSESSING THE RELIABILITY OF THE FOUR AWARENESS CRITERIA

Reliability Between Criteria

The original Olson and Fazio (2001) criterion used in Moran et al.'s (2021) primary analysis was the only exclusion criterion under which a significant EC effect was found. As outlined above, this criterion was also the most liberal one. While the awareness rates produced under different criteria were reported by Moran et al., that article did not address the relationship between relative strictness of the criteria and the EC effects that they produced.

The question of whether observed differences in exclusion rates across exclusion criteria could be attributed to differences in their strictness (a desirable property) versus their unreliability or poor measurement (an undesirable property) is testable (Guttman, 1944; Meijer, 1994). Such tests can be conducted by considering a statistical property known as the degree of conformity to a Guttman structure, which is estimable using methods from Item Response Theory (IRT) modeling. Specifically, if these measures as a set demonstrated perfect reliability and differed only in their strictness, we would expect the proportion of Guttman errors (G) to be very small (i.e., approach 0). In contrast, if the differences between them were attributable exclusively to their unreliability, we would expect G to approach 1.

In the context of the awareness criteria, there were observed differences in exclusion rates in the sample as a whole. Specifically, under the original Olson and Fazio (2001) criterion 7.6%, under the modified Olson and Fazio (2001) criterion 30.6%, under the original Bar-Anan et al. (2010) criterion 47.9%, and under the modified Bar-Anan et al. (2010) criterion 26.9% of participants were excluded from analyses. Under the assumption of perfect measurement properties, this pattern of results would be entirely due to the measures differing in their relative strictness rather than lack of reliability between them.

As such, we sought to examine whether the differences across exclusion criteria were due to differences in the 'difficulty' of the items (i.e., their location along the continuum that is better referred to as 'strictness' in this case). Specifically, if the items are collectively reliable, the individuals who were scored as 'aware' on a criterion that excluded the lowest proportion of the sample should also be scored as 'aware' on a criterion that excluded the highest proportion of the sample. If not, then the criteria, as a set, did not function as reliable measures of awareness in the first place. This approach can be understood using an analogy with aptitude testing. On a good test, the 'easy' questions are those that most individuals get correct, and the 'difficult' questions are those that few individuals get correct. As such, if someone gets a 'difficult' question correct, they should also have gotten the questions that were relatively easier correct (for a more technical discussion, see Guttman, 1944).

Results from the IRT analysis suggested that the awareness measures were quite unreliable. Nearly half of participants had scores on one or more awareness criteria that indicated Guttman errors, G = 47.5%, 95% CI [45.5, 49.5], G* = 11.9%, 95% CI [11.4, 12.4] (where G* = G/(items – 1); see Meijer, 1994, for a discussion of G and its standardized form G*). In other words, about half of the participants were scored as aware by an item with a relatively low exclusion rate while also being scored as unaware by an item with a relatively high exclusion rate. Overall, this pattern of results suggests that differences across exclusion criteria are, to a large degree, attributable to unreliability of the measures.[1]

Heterogeneity Between Sites

Given that all measures and instructions were delivered to participants in a standardized format, a large degree of heterogeneity in rates of contingency awareness across data collection sites may imply that the awareness measures are not as valid (or uniformly valid across sites) as assumed. Indeed, we found considerable variation in exclusion rates at the site level: For example, exclusion rates using the Olson and Fazio (2001) modified criterion varied between 15% and 74%. Such variability can be quantified using meta-analyses of the proportion of aware participants between sites for each of the exclusion criteria. Results demonstrated large between-site heterogeneity for all four criteria (all I^2 = 54.7% to 91.7%, all H^2 = 2.2 to 12). Differences in between-site awareness rates therefore did not represent mere sampling variation but rather large between-site heterogeneity. Such heterogeneity could be attributed to the somewhat subjective nature of the Olson and Fazio (2001) criterion in particular, which (*a*) asks participants the broad question of whether they 'noticed anything odd during the experiment,' (*b*) collects open-ended responses, and (*c*) requires these responses to be hand scored. Alternatively, the differences could represent genuine differences in awareness rates across sites; however, if this is the case, this finding raises the question of why such stark (up to five-fold) differences may have emerged although the sites did not substantively differ from each other culturally.

EC UNDER A STRICTER COMPOUND CRITERION

Considering that the possibility of EC in the presence of awareness is uncontroversial, the conclusion that EC can

emerge in the absence of awareness requires a severe test of the hypothesis. Such a test presupposes making the maximum effort to exclude participants who are aware of the stimulus pairings. As such, we created a stricter exclusion criterion that maximized the chances of excluding aware participants by prioritizing sensitivity over specificity. We believe that the specificity of the awareness measure is relatively unimportant in the present context: Incorrectly excluding some unaware participants from the analysis seems acceptable as long as (*a*) all efforts are made to exclude aware participants and (*b*) the remaining sample provides sufficient power to test the hypothesis.

Specifically, we excluded participants if one or more of the four awareness criteria scored them as aware of the CS–US pairings. This compound criterion excluded 54% of participants as aware, leaving 665 in the analytic sample. In this subsample, using the power analysis method employed by Moran et al. (2021), power to detect an effect size as large as that observed in the published literature (i.e., $g = 0.20$) was >.99. Power estimates were comparable when we employed what we considered to be a more appropriate method of power analysis for meta-analytic models (Valentine et al., 2009): to detect an effect size of $d = 0.20$, power was = .95. As such, the available sample size provided adequate statistical power for the analysis reported below, comparable to Moran et al. (2021).

After excluding participants using the compound criterion, we fit a meta-analytic model that was otherwise identical to that used in Moran et al.'s (2021) primary analysis. The meta-analyzed EC effect was a non-significant, well-estimated effect size that was exceptionally close to zero, Hedges' $g = 0.00$, 95% CI [–0.11, 0.10], $p = .983$. No heterogeneity was observed between sites, $I^2 = 0.0\%$, $H^2 = 1.0$ (see Figure 1).

A Bayes Factor meta-analytic model using Rouder and Morey's (2011) method was also fit to quantify the evidence in favor of the null hypothesis. Default JZS and Cauchy priors were employed to represent a weak skeptical belief in the null hypothesis (location = 0; scaling factor $r = .707$ on the fixed effect for condition and $r = 1.0$ on the random effect for data collection site, see Rouder & Morey 2011). This analysis provided strong evidence in favor of the null hypothesis ($BF_{01} = 22.83$, effect size $\delta = 0.00$, 95% HDI [–0.08, 0.07]).

A reviewer of this work suggested that (*a*) as Moran et al.'s uncorrected meta-analysis of published results yielded an effect size of 0.20, 0.20 is a reasonable maximum effect size to be considered, and (*b*) 95% of a Cauchy distribution lies within 7 scaling factors. As such, the reviewer argued that a more appropriate scaling factor on the fixed effects would be 0.20/7. We calculated a new meta-analysis using this scaling factor rather than the default. Results from this new, exploratory analysis suggested no strong evidence for either the null or the alternative hypothesis ($BF_{01} = 1.64$). Nonetheless, the effect size was estimated even more precisely than before as close to zero (effect size $\delta = 0.00$, 95% HDI [–0.05, 0.05]).

VALIDITY OF STATISTICAL INFERENCES I: BAYESIAN ANALYSES

Setting aside questions about the validity of the exclusion criteria, the present Bayesian analyses gave us the opportunity to reexamine the robustness of multiple focal statistical inferences emerging from the replication project by Moran et al. (2021). Specifically, we investigated three separate, but interrelated, questions: (*a*) Do Bayesian analyses provide compelling evidence

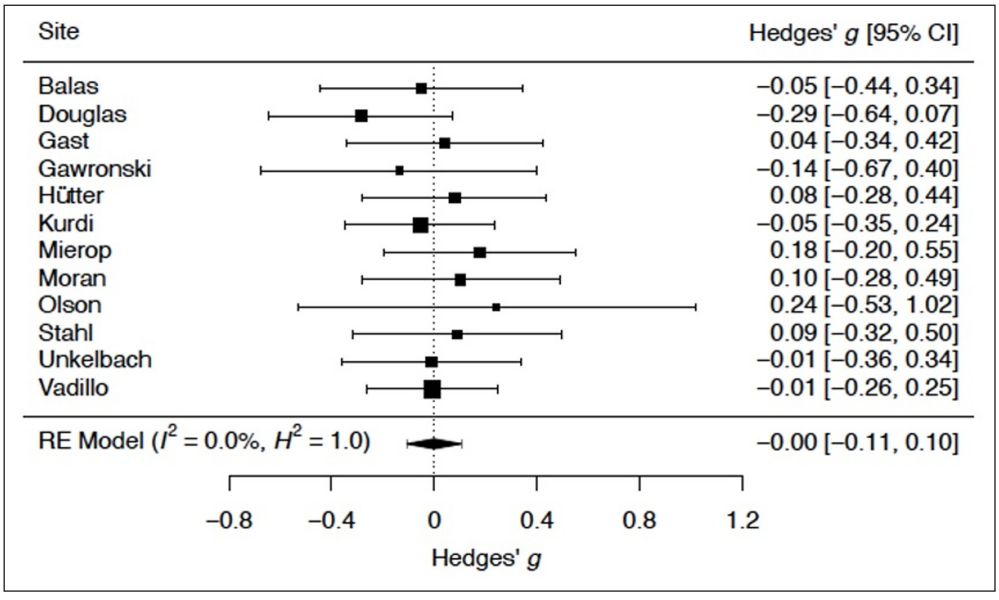

Figure 1 Forest plot of the meta-analytic results as a function of data collection site.

for an unaware EC effect under the original Olson and Fazio (2001) exclusion criterion? (*b*) Are inferences about the presence of an unaware EC effect robust to exclusion criteria? (*c*) Are the inferences about the presence of an unaware EC effect robust to whether the analysis considers only data by Moran et al. (2021) or data from the surveillance task literature as a whole? To this end, we repeated our analyses under three different choices of prior on the estimate of the unaware EC effect. These priors included (*i*) an uninformative (default) prior, which allowed us to estimate the effect that emerged specifically in the context of the Moran et al. (2021) study, without considering any extraneous information, (*ii*) an informative prior relying on the unadjusted meta-analytic estimate of the effect size, and (*iii*) an informative prior relying on the meta-analytic estimate adjusted for publication bias.

Beyond the ability to incorporate, and explicitly compare the effects of, multiple reasonable priors, Bayesian analyses also offer other benefits over the frequentist analyses reported by Moran et al. (2021). Notably, with large samples, frequentist analyses can yield statistically significant results even when the null hypothesis is more likely to be true than the alternative hypothesis. As such, it is conceivable that Bayesian analyses may find evidence for the null hypothesis, or uncover a considerable degree of uncertainty, even when a frequentist statistical test is 'statistically significant' by conventional standards, as it was in the Moran et al. (2021) paper under the original Olson and Fazio (2001) exclusion criterion.

Another potential drawback of Moran et al.'s approach is that non-significant findings are inherently ambiguous in a frequentist framework (Dienes, 2014). Specifically, they could generally indicate either lack of adequate power or the genuine absence of an effect. In the specific context of the surveillance task, it is possible that the meta-analytic tests of moderation by inclusion criteria were not sufficiently well-powered to detect even large differences. Alternatively, if adequately powered, they could have indicated that the effects obtained under different exclusion criteria truly do not differ from each other in the population. As such, it is important to probe whether participant exclusion criteria other than that used by Olson and Fazio (2001) provide evidence in favor of the null hypothesis, and whether different exclusion criteria yield qualitatively similar or different results.

To conduct the Bayesian analyses, we fit intercept-only mixed effects models to the data, with standardized preference for the CS_{pos} over the CS_{neg} as the dependent variable and random intercepts for data collection sites. In model 1, we placed a default (uninformative) prior on the intercept; in model 2, we used the unadjusted meta-analytic effect size reported by Moran et al. (2021) as an informative prior; and in model 3, we used the publication bias-adjusted meta-analytic effect size for the same purpose. As mentioned above, such variation in priors (ranging from noncommittal to quite optimistic to quite skeptical) provides some indication about the robustness of the results emerging from any individual study. Moreover, the informative priors based on the meta-analysis reported by Moran et al. (2021) allow us to characterize the strength of unaware EC effects emerging from the surveillance task literature as a whole rather than from the replication study alone.

We preregistered the standardized regression coefficient β = 0.10 as the smallest effect consistent with the directional alternative hypothesis (H_1); the area of the posterior distribution below this value was seen as consistent with the null hypothesis (H_0). This threshold was chosen a priori because it is widely used in Bayesian equivalence testing (Kruschke, 2018) given that it represents half of what is usually considered to be a small effect. Specifically, a mean difference below this value would indicate that evaluations of the CS_{pos} and CS_{neg} differ from each other by less than one tenth of a standard deviation. We note that this threshold is fairly liberal considering that even arguably diminishingly small effects are considered to be consistent with the alternative hypothesis. Moreover, investigators who wish to consider the posterior distributions in their totality can do so by examining them in Figure 2.

The main quantity of interest, upon which we base statistical inferences, is the proportion of the posterior distribution consistent with H_1 versus H_0. This is a continuous quantity that represents varying degrees of evidence in favor of H_1 or H_0, respectively. Specifically, if half the posterior distribution were below $\beta = 0.10$ and the other half were above $\beta = 0.10$, then (depending on one's prior expectations) the data may be seen as completely uninformative given that H_1 or H_0 would be equally likely to be true. Any deviation from equiprobability can be seen as providing some level of support for H_1 or H_0. However, if more than 95% of the posterior are found to be consistent with H_1 or H_0, we refer to this as decisive evidence in favor of the corresponding hypothesis (e.g., Kruschke, 2018).

EC EFFECT UNDER THE ORIGINAL OLSON AND FAZIO (2001) EXCLUSION CRITERION

The outcome of the Bayesian reanalysis of the Moran et al. data using default priors is shown in the top row of Figure 2. This reanalysis makes it clear that the original Olson and Fazio exclusion criterion provides the relatively strongest evidence in favor of unaware EC effects. However, with over 23% of the posterior distribution favoring H_0, the evidence for H_1 is not decisive. As such, the present analysis strongly qualifies the main conclusion of the replication project: Although a statistically significant result may have been obtained in the original analysis, the Bayesian reanalysis shows that considerable portions of the posterior distribution around the estimate are

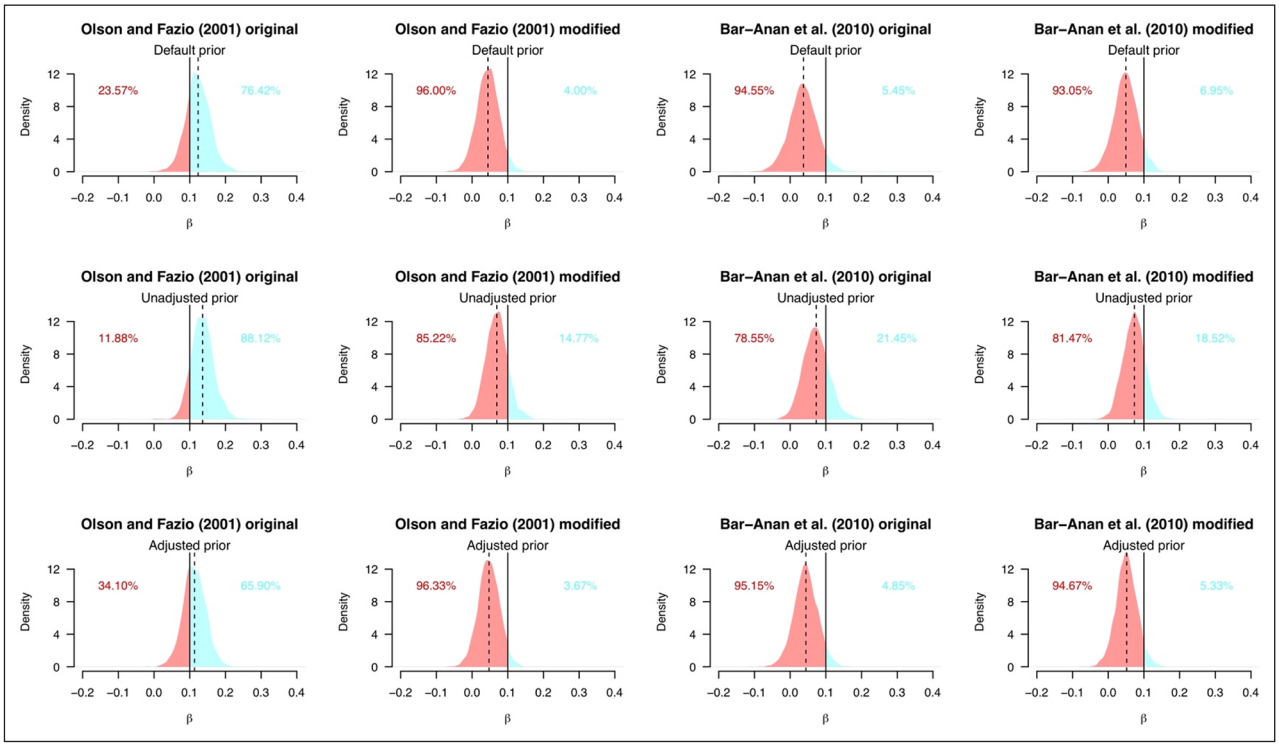

Figure 2 Posterior distribution of the unaware EC effect under different priors and exclusion criteria. Positive scores correspond to the theoretically expected preference for CS_{pos} over CS_{neg}. The dashed vertical line shows the posterior mean, and the solid vertical line shows the smallest effect size of interest ($\beta = 0.1$). Areas displayed in red are consistent with the null hypothesis H_0: $B < 0.1$ and areas displayed in light blue are consistent with the directional alternative hypothesis H_1: $B \geq 0.1$. Percentages denote the proportion of the posterior distribution consistent with H_0 and H_1, respectively.

consistent with a minuscule, or even negative, effect. This result is remarkable in its ambiguity: The Moran et al. (2021) study relied on an extremely large sample of over 1,400 participants and even this extremely large sample was insufficient to produce clear evidence for an effect.

EC EFFECT UNDER ALTERNATIVE EXCLUSION CRITERIA

When alternative exclusion criteria were used, the data provided convincing evidence for H_0, with proportions of the posterior distribution favoring no effect falling between 93% and 96%. As such, unlike the original frequentist analysis, the current Bayesian analysis warrants the conclusion that the non-significant results obtained using alternative exclusion criteria did not emerge due to lack of statistical power (perhaps due to the increasingly small portions of the sample being considered unaware under increasingly conservative criteria); rather, the present models positively suggest an absence of unaware EC effects. The present analyses also indicate that, in addition to the issues of dependence addressed in the section below, the lack of significant difference between exclusion criteria was most likely due to inadequate statistical power: Comparing different exclusion criteria in a Bayesian framework (top row, Figure 2) makes it clear that the original exclusion criterion and alternative exclusion criteria result in qualitatively different conclusions from the same data. Specifically, whereas the former yields some limited support for H_1 over H_0, the latter provide strong evidence in favor of H_0 over H_1.

EC EFFECT UNDER INFORMATIVE PRIORS

Finally, we fit the same models using the unadjusted meta-analytic effect size (row 2, Figure 2) and the adjusted meta-analytic effect size (row 3, Figure 2) as informative priors, thus explicitly incorporating the results of previous work relying on the surveillance task into the analyses. The most important takeaway from these analyses is that, presumably given the large sample used in the replication project, the results seem quite robust to the choice of prior. Specifically, even the most lenient analysis using the unadjusted meta-analytic effect size as the informative prior and the original exclusion criterion does not provide unequivocal evidence for H_1 over H_0: Close to 12 percent of the posterior remained consistent with the null hypothesis. Under alternative exclusion criteria, we found robust evidence for H_0 over H_1, including under the arguably overly optimistic assumption of no publication bias in the surveillance task literature.

To summarize, these Bayesian analyses cast considerable doubt on the possibility of unaware EC effects in the surveillance task. Specifically, in an analysis relying on uninformative priors and the original Olson and Fazio (2001) exclusion criterion, we found that a considerable portion of the posterior distribution was consistent with the null hypothesis, thus questioning

whether the replication attempt was an 'unqualified' success. Second, we obtained compelling evidence *against* unaware EC effects using alternative (and, arguably, more appropriate) exclusion criteria. Finally, these inferences were relatively robust to the choice of prior, including an informed prior relying on meta-analytic estimates of the effect size. As such, the surveillance task literature as a whole does not seem to provide convincing evidence in favor of unaware EC effects.

VALIDITY OF STATISTICAL INFERENCES II: CLASSIFICATION ANALYSIS

Although the Bayesian analyses reported above substantially qualify the conclusions that can be drawn from the data obtained by Moran et al. (2021), they do not address one crucial shortcoming of the analyses reported in that paper: namely, the dependency between and discrepancy across different criteria of contingency awareness. It is to this issue that we turn in the final empirical section of the present paper, using a classification approach.

Similar to the previous section, we set aside concerns about the validity of the correlational approach to unaware EC in general and about the validity and reliability of the measures used by Olson and Fazio (2001) in particular. We also set aside concerns about the statistical analysis that have been addressed in the previous section. Instead, we focus on whether the data produced by Moran et al. (2021) provide evidence for truly unaware conditioning effects, under the assumption that the awareness measure is valid and that the frequentist analytical approach is well-suited to the task at hand. Crucially, in doing so, we revisit the issue of whether—and why—different exclusion criteria produce significantly different results.

The analytic rationale in the original Olson and Fazio (2001) study and the RRR was to retain only ostensibly unaware participants for statistical analyses. If an EC effect is found in the unaware subsample, then this result is then taken as evidence for unaware EC. Notably, conclusions about unaware EC crucially depended on the choice of criterion used to exclude participants from subsequent analyses: Evidence for unaware EC was obtained only under the original criterion used by Olson and Fazio (2001), but not under any of the other criteria. This result suggests that the choice of criterion matters, and that different criteria yield different conclusions. Yet, Moran et al. (2021) also reported that the EC effects obtained under the four different criteria did not significantly differ from each other. This latter result suggests that the choice of criteria does not matter, which would imply that they should point to the same conclusion. As mentioned above, this set of findings is ambiguous with regard to the question of unaware EC—do the results support or oppose its existence?

Here we show that these contradictions are easily resolved when identifying a new group of partially aware participants. In doing so, we argue that the question of moderation across criteria does not address the issue of unaware EC (and should be disregarded), and that the apparent evidence for unaware EC obtained by the original criterion was caused by the inclusion of partially aware participants in the unaware category.

As explained above, Moran et al. (2021) considered two pairs of exclusion criteria for classifying participants as 'aware.' Here we focus on the first pair of criteria, which have resulted in qualitatively different results in Moran et al. (2021) as well as in the previous section: The relatively liberal original Olson and Fazio (2001) criterion (which produced evidence for unaware EC) and a more stringent modified version of this criterion (which did not). Despite the fact that these two criteria show qualitatively distinct results regarding the presence of unaware EC, their estimates of unaware EC do not differ significantly from one another.

A nonsignificant moderation of EC by criterion, as reported by Moran et al. (2021) for all four criteria, is also obtained when focusing on only the first two criteria: In line with Moran et al.'s conclusions, the EC effect estimated under the original (g = .12) and modified (g = .05) criteria did not differ significantly, $QM_{(df = 1)} = 1.67$, $p = 0.196$. This test, however, is misleading, because it compares two dependent, largely overlapping samples: The data from participants classified as 'unaware' by the modified criterion are included under both criteria.

As discussed above, these two criteria are based on the same data (i.e., participants' responses to two open-ended questions) and differ only in how these data were coded: While the original criterion classified cases of partial awareness as 'unaware,' the modified criterion classified partial-awareness cases as 'aware.' Cases of partial awareness included participants mentioning only the CS–US pairings of one valence; participants misreporting which CS was paired with which US; and participants referring to systematic CS–US pairings without specifying the valence of the USs with which each CS was paired. We believe that these participants should be classified as 'aware' because the information that they reported was sufficient to produce a conscious EC effect.

Moran et al. (2021) found an EC effect only when using the more liberal original criterion (which included unaware as well as partially aware participants in the 'unaware' category); the effect disappeared when using the more stringent modified criterion (which included only unaware participants and excluded partially aware cases). This pattern of results implies that the inclusion of partially aware participants in the unaware category drove the 'unaware' EC effect obtained under the original

criterion, because unaware EC was no longer significant when partially aware participants were classified as 'aware' under the modified criterion.

In addition, the current discussion highlights why a comparison of EC effects across criteria is neither methodologically sound (because it compares overlapping samples, namely the unaware and partially aware participants, taken together, are compared to the unaware participants), nor helpful in answering the question of whether there is, in fact, an EC effect among unaware participants. Instead, one should rely on the better one of the two estimates of unaware EC, that is, the one that excludes partially aware participants from the unaware category.

THREE—NOT TWO—SUBGROUPS NEED TO BE DISTINGUISHED: FULLY AWARE, PARTIALLY AWARE, AND UNAWARE PARTICIPANTS

Reflecting the differences in stringency discussed above, the two exclusion criteria differed in how many participants they excluded as aware: Of the total N = 1,450, the original authors' criterion excluded 8% (n = 110), whereas the modified criterion excluded 31% (n = 443). That is, in addition to the 'aware' participants according to both criteria, there were 23% of 'partially aware' participants who were classified as 'unaware' by the original but not by the modified criterion. In other words, the two criteria, considered jointly, yield three subgroups: (*a*) a small fully aware subgroup (n = 110, 8% of the data; 'aware' by both criteria), (*b*) a medium-sized partially aware subgroup (n = 333, 23%; 'aware' only by the modified criterion but 'unaware' by the original criterion), and (*c*) a large fully unaware subgroup (n = 1,007, 69%; 'unaware' by both criteria).

Note again that the original authors' criterion sets a lower bar for classifying participants as 'unaware,' leaving partially aware participants in the unaware category. Therefore, EC effects associated with partial awareness are interpreted as unaware, and the EC-without-awareness test is easier to pass. Using this original criterion, the replication yielded a significant EC effect. In contrast, the bar for an 'unaware' classification according to the modified exclusion criterion is more conservative, excluding partially aware participants from that category. The unaware EC effect for this criterion was therefore unaffected by partial awareness, yielding a more stringent test of the EC-without-awareness hypothesis.

We already saw in Moran et al. (2021) that the replication data failed to pass this more stringent test: There was no evidence for EC when the modified Olson and Fazio (2001) criterion was used. Furthermore, the Bayesian analyses reported in the previous section obtained evidence for the absence of EC with that criterion. Given that (*a*) the original criterion is contaminated by 'partially aware' participants and (*b*) the modified criterion considers only the relevant subgroup, the latter is clearly the more appropriate one to use. Separate EC estimates for the three subgroups illustrate this point and extend the Moran et al. (2021) results.

EC EFFECTS AMONG FULLY AWARE, PARTIALLY AWARE, AND UNAWARE PARTICIPANTS

We used the subgroup classification established above as a moderator in a random-effects meta-analysis that investigated EC effects across the three subsamples. Crucially, results confirmed the considerations explained above (see Figure 3): Subgroup membership moderated the magnitude of the EC effect, $Q_{(df = 2)} = 12.39$, $p = .002$. The EC effects for partially aware (g = .36, 95% CI [.19, .53]) and fully aware participants (g = .31, 95% CI [.02, .59]) were of comparable (small-to-medium) magnitude and significantly different from zero, despite representing a relatively small subset of the data. When only unaware participants were considered, evidence for EC was absent (see Figure 3): In this subgroup, the EC effect was very small and its confidence interval contained zero, (g = .05, 95% CI [–.04, .13]). Unsurprisingly, this result is exactly the 'unaware EC' estimate reported under the modified criterion by Moran et al. (2021).

As such, the apparent contradictions are logically and empirically resolved: Discrepant conclusions about unaware EC using different criteria and ambiguous conclusions derived from comparing different criteria to test unaware EC are now shown to be driven by the inclusion of a partially aware group of participants in the unaware category when using the original criterion. In turn, this result calls into question the original authors' conclusion of an 'unqualified replication' of the Olson and Fazio (2001) finding. As the present analysis indicates, this 'successful replication' depends on the misclassification of a subgroup of participants as 'unaware' who, in all likelihood, were (partially) aware of the CS–US pairings.

GENERAL DISCUSSION

Science is a self-correcting process, and RRRs are a useful tool to that end. However, when large groups of collaborators—including the current authors, when it comes to Moran et al. (2021)—are involved in that process, compromises become necessary, and these may undermine the clarity of the conclusions drawn. Thus, with the spirit of self-correction in mind, we revisited the data from the Moran et al. study. Below, we offer some alternative conclusions about unaware EC that deviate from those presented in the replication report. In closing, we highlight a more general issue when replicating classic experiments in psychological science.

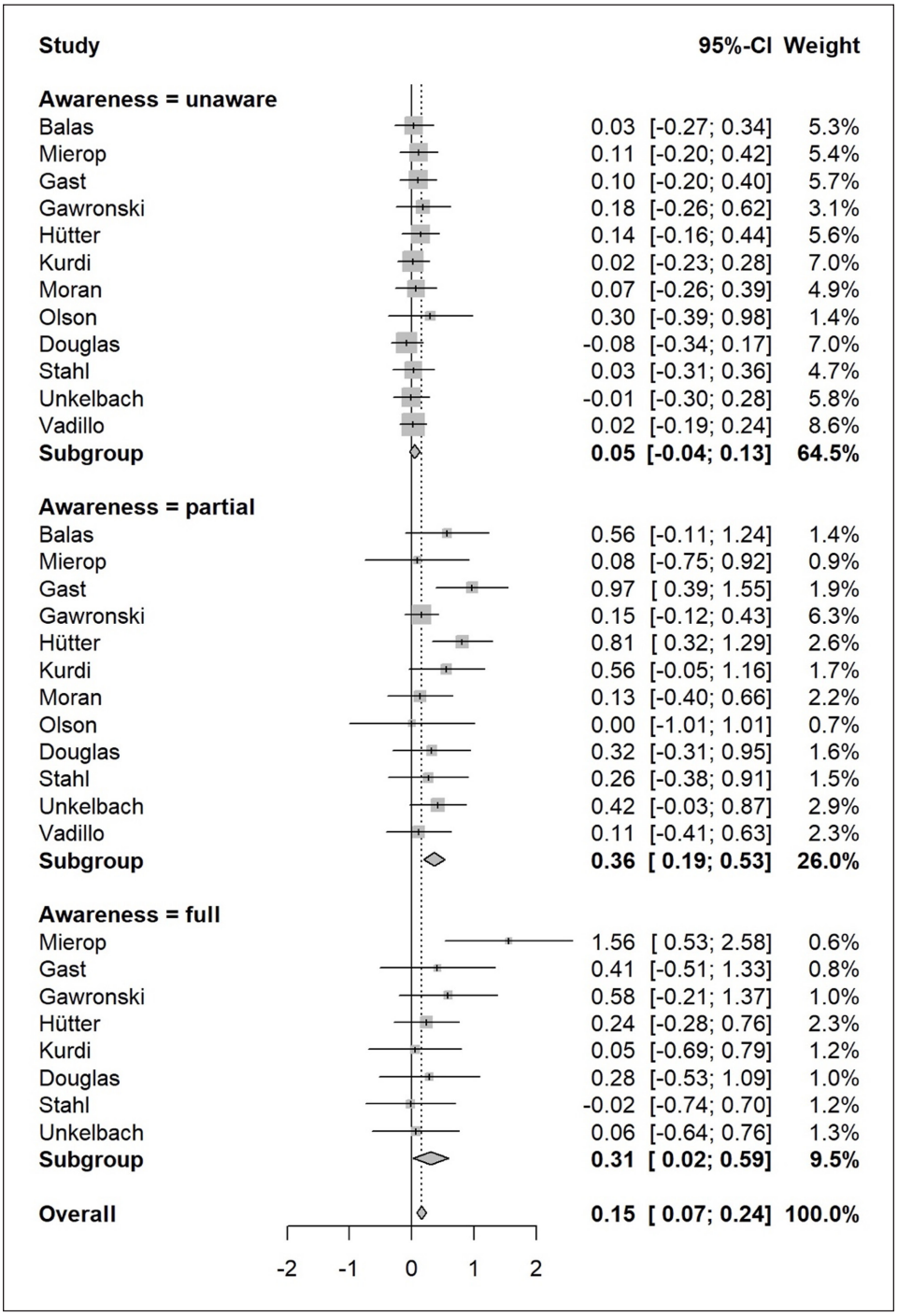

Figure 3 Meta-analytic results as a function of the awareness sub-type criterion. There were four sites with only two fully aware participants for which the corresponding effect sizes could not be calculated.

DATA BY MORAN ET AL. (2021) PROVIDE EVIDENCE AGAINST UNAWARE EC

As mentioned above, reflecting on the Moran et al. findings, the original authors concluded the following: 'Ultimately, the lack of a moderating effect of exclusion criteria can be interpreted as an unqualified replication of Olson and Fazio (2001)' (p. 129). That is, they considered the replication of an EC effect under their original exclusion criterion as providing unequivocal support for the idea that the surveillance task provides an accurate, reliable, and useful measure of unaware EC. We respectfully disagree.

The novel analyses of the Moran et al. (2021) data that we report here question the validity of the awareness measures used in the surveillance task studies. If anything, the present analyses indicate that the awareness criteria used in both the original Olson and Fazio (2001) study as well as the replication by Moran et al. (2021) are relatively poor measures that likely fail to exclude genuinely aware participants. When subjected to a more severe test that prioritized sensitivity, Moran et al.'s (2021) data did not support the unaware EC hypothesis. This finding highlights the importance of distinguishing between a replicable statistical effect (which was found by Moran

et al.) and a replicable inference regarding a hypothesis of interest (which was not found here; for more on this distinction see Vazire, 2019; Yarkoni, 2020).

Putting considerations regarding the psychometric qualities of the awareness measures aside, in a Bayesian analysis, we did not find clear evidence for unaware EC effects even under the most liberal exclusion criterion and the most optimistic choice of priors. In fact, arguably more reasonable exclusion criteria and less overly optimistic prior specifications provided compelling evidence *against* the idea of unaware EC in the surveillance task. In addition, in a classification analysis, EC effects were moderated by awareness type, thus undermining the original conclusion that different exclusion criteria produced equivalent results. Instead, it seems that the original findings of 'unaware' EC were due to a subgroup of aware participants erroneously classified as unaware given the open-ended and ambiguous nature of the Olson and Fazio (2001) awareness measure.

Specifically, we obtained no evidence that EC effects emerged in fully unaware individuals, although a small effect was observed among those with partial or complete awareness of the CS–US contingency. Of course, it is conceivable that alternative measures of awareness (or evaluation) may have identified a small subset of participants who are unaware of the contingencies and yet show significant EC effects (for related results, see Jurchis et al., 2020; Waroquier et al., 2020). However, crucially, even if this were the case, any correlational measure has only limited ability to inform about awareness during encoding of the CS–US pairings. We return to this point in more detail below.

CONCEPTUAL ISSUES SURROUNDING AWARENESS, MEMORY, AND INTENTION

Irrespective of the particular results obtained in the reanalyses reported above, in closing it seems worth revisiting some of the conceptual ambiguities inherent in this work. Notably, beyond disagreements considering the strength of evidence for EC effects in the relevant subgroup of participants, the authors of Moran et al. (2021) also disagreed on what exactly it was that the retrospective self-report measure used to delineate that relevant subgroup captured. Some interpreted it as a measure of a specific type of (contingency) awareness, whereas others interpreted it as a measure of recollective (contingency) memory. Specifically, in trying to integrate the various perspectives of its contributing authors, the RRR navigated between three notions: (*a*) the incidental nature of the surveillance task, (*b*) unaware attitude formation, and (*c*) the role of contingency memory in EC effects.

However, as already alluded to in the Introduction, these notions should not be confounded with each other. The incidental nature of the surveillance task refers to the absence of intentional attitude formation. Whether EC effects emerge unintentionally is an important question in its own right (Stahl et al., 2016). However, intentionality does not perfectly align with another feature of automaticity: awareness (e.g., Moors & De Houwer, 2006). In addition, no measure of intention was collected in the current studies, and therefore we do not know which participants may or may not have intentionally formed an attitude during the learning phase of the experiment.

Admittedly, the surveillance task makes it more difficult for participants to consciously encode the CS–US pairings and it does not heavily direct attention toward the evaluative implications of these pairings. It is therefore interesting to probe what information is encoded in memory in this paradigm, and how participants' capacity to retrieve this information relates to EC effects. However, as extensively discussed in previous work (e.g., Corneille & Stahl, 2019; Gawronski & Walther, 2012; Sweldens et al., 2014), examining this relation in the context of a purely correlational design allows only for conclusions about whether EC effects can be obtained in the absence of memory at the evaluation stage. Correlational designs do not warrant conclusions about whether attitudes were formed in the absence of awareness at encoding.

Investigations of unaware attitude formation require experimental manipulations of awareness during encoding. Whether the surveillance task can be used to achieve this goal is ambiguous. Truly experimental studies that manipulated awareness at encoding have largely failed to support unaware attitude formation in EC procedures (for a review, see Corneille & Stahl, 2019). Therefore, the question that Moran et al.'s RRR addressed was whether, in a procedure that can reasonably be considered incidental, EC effects can emerge in the absence of retrospective memory for the CS–US pairings. This is a valuable question, but one that does not necessarily speak to unconscious or unaware attitude formation.

Finally, even if it were possible to resolve these conceptual issues, and draw firm conclusions about the underlying process, it seems important to keep in mind that the retrospective measure used by Olson and Fazio (2001) does not appear to be sufficiently valid or sensitive. Rather, it seems to index the non-cued retrieval of information stored in memory, among participants who may or may not have been motivated to do their best in retrieving that information. A more valid and sensitive measure is needed to accurately detect contingency memory and may also require incentives to be delivered to motivate accurate recall.

SOME CLOSING REMARKS ON THE VALUE OF REPLICATIONS

The experiments by Olson and Fazio (2001) are classic entries into the canon of research on evaluative learning and, at the time of their publication, they provided

intriguing evidence for EC without awareness. Replicating these findings was an inherently valuable undertaking that informs the question of unaware or implicit attitude formation. And, as all of the present authors can attest, the lead authors of the Moran et al. (2021) RRR project went to great lengths to conduct a fair replication of the original experiments.

The additional analyses and considerations discussed here highlight another aspect of replication studies. The Moran et al. (2021) paper includes a coda by the original authors claiming that the RRR presents an unqualified replication of Olson and Fazio (2001). It may have been wise to add: 'using the methods of 2001.' With this statement, we are in full agreement; this is what the data show. However, science is a progressive endeavor, and when it comes to conceptual considerations (e.g., distinctions between incidental vs. unaware aspects of the learning task), methodological considerations (e.g., how to probe for CS–US memory), and statistical considerations (e.g., advanced meta-analytical tools and more widespread use of Bayesian techniques), EC research has evolved substantially beyond the state of the art twenty years ago.

When the data are evaluated according to contemporary standards, there seems to be little evidence to suggest that EC effects can emerge in the absence of awareness. The same way a celestial object (here: awareness) might go undetected because the telescope is too small, a larger telescope or a better scanning program may provide clear evidence for the searched object (i.e., awareness among participants who show an EC effect). As such, the RRR by Moran et al. (2021) provided solid evidence that Olson and Fazio's (2001) conclusions were correct using the best methods available at the time. However, the present analyses indicate that these conclusions do not hold when one probes for awareness using improved measures or analytic strategies.

In closing, we note that we are not brushing aside the possibility of implicit learning, or even that of unconscious attitude formation. Given evidence for incidental learning effects across several domains of human cognition (e.g., Knowlton et al., 1992; Saffran & Kirkham, 2018; Schapiro et al., 2013), it is plausible that, under certain conditions, EC effects may emerge in the absence of contingency awareness. Nevertheless, the surveillance task is not a paradigm that is well-equipped to speak to this issue.

More generally, as discussed above, studies conducted using alternative, and arguably more rigorous, approaches have so far also failed to produce convincing evidence for unaware EC effects. In addition, more parsimonious single-process approaches to learning and memory are often able to accommodate findings that are usually considered to support dual-process models. However, given the theoretical importance of the issue, we do not anticipate that the search for robust and replicable demonstrations of unaware evaluative learning will be abandoned anytime soon.

DATA ACCESSIBILITY STATEMENT

The data collected as part of the RRR by Moran et al. (2021) are available for download from the Open Science Framework (OSF; http://osf.io/hs32y).

The analyses reported under 'Validity of Statistical Inferences I: Bayesian Analyses' were preregistered (https://osf.io/sfh4v/). The remaining analyses were not formally preregistered.

Computer code for reproducing the analyses reported in the present paper is available on OSF, under https://osf.io/ugrjh/ for 'Validity of the Olson and Fazio (2001) Contingency Awareness Measures'; under https://osf.io/u4mv8/ for 'Validity of Statistical Inferences I: Bayesian Analyses'; and under https://osf.io/qs35v/ for 'Validity of Statistical Inferences II: Classification Analysis.'

The analyses were conducted in the R statistical computing environment using the R packages BayesFactor (version 0.9.12–4.2; Morey & Rouder, 2018), brms (version 2.14.4; Bürkner, 2017), dplyr (version 0.8.4; Wickham et al., 2020), meta (version 4.11.0; Balduzzi et al., 2019), metafor (version 2.4–0; Viechtbauer, 2010), and papaja (version 0.1.0.9942; Aust & Barth, 2020).

NOTE

1 In response to a reviewer comment, we ascertained that the large percentage of Guttman errors was not the function of any one exclusion criterion: The estimate of Guttman errors ranged from 32.4% to 46.6% in four subset analyses omitting one exclusion criterion at a time.

FUNDING INFORMATION

Preparation of this work was supported by FRS-FNRS grant #T.0061.18 (to O. Corneille).

COMPETING INTERESTS

B. Kurdi is a member of the Scientific Advisory Board at Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition.

AUTHOR CONTRIBUTIONS

B. Kurdi, I. Hussey, and C. Stahl were first authors of the three commentaries that form the basis of this article and thus share co-first authorship.

I. Hussey conceptualized and conducted the data analysis reported in the section 'Validity of the Olson and Fazio (2001) Contingency Awareness Measures'; S. Hughes contributed to the design and data analysis; I. Hussey and S. Hughes contributed to the writing and revision of the article.

B. Kurdi and M. J. Ferguson conceptualized the data analysis reported in section 'Validity of Statistical Inferences I: Bayesian Analyses'; B. Kurdi conducted data analysis; B. Kurdi drafted the section; M. J. Ferguson contributed to the review of the section; B. Kurdi and M. J. Ferguson contributed to the writing and revision of the article. B. Kurdi coordinated revisions of the manuscript.

C. Stahl conceptualized and conducted the data analysis reported in the section 'Validity of Statistical Inferences II: Classification Analysis'; C. Stahl and O. Corneille co-wrote the section; C. Unkelbach contributed to the review of the section; C. Stahl, O. Corneille, and C. Unkelbach contributed to the writing and revision of the article. O. Corneille drafted and helped revise the Introduction and General Discussion sections.

All authors approved the final version of the manuscript for submission.

AUTHOR AFFILIATIONS

Benedek Kurdi orcid.org/0000-0001-5000-0584
Yale University, US

Ian Hussey orcid.org/0000-0001-8906-7559
Ruhr University Bochum, DE

Christoph Stahl orcid.org/0000-0002-9033-894X
University of Cologne, DE

Sean Hughes orcid.org/0000-0001-7689-4272
Ghent University, BE

Christian Unkelbach orcid.org/0000-0002-3793-6246
University of Cologne, DE

Melissa J. Ferguson orcid.org/0000-0003-2840-0033
Yale University, US

Olivier Corneille orcid.org/0000-0003-4005-4372
UCLouvain, BE

REFERENCES

Aust, F., & **Barth, M.** (2020). *papaja: Create APA manuscripts with RMarkdown*. R package version 0.1.0.9942. https://www.rdocumentation.org/packages/papaja/versions/0.1.0.9942

Baeyens, F., Vansteenwegen, D., & **Hermans, D.** (2009). Associative learning requires associations, not propositions. *Behavioral and Brain Sciences*, *32*(2), 217–218. DOI: https://doi.org/10.1017/S0140525X09000867

Balduzzi, S., Rücker, G., & **Schwarzer, G.** (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence Based Mental Health*, *22*(4), 153–160. DOI: https://doi.org/10.1136/ebmental-2019-300117

Bar-Anan, Y., De Houwer, J., & **Nosek, B. A.** (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *Quarterly Journal of Experimental Psychology*, *63*(12), 2313–2335. DOI: https://doi.org/10.1080/17470211003802442

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

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: https://doi.org/10.1177/1088868318763261

Davies, S. R., El-Deredy, W., Zandstra, E. H., & **Blanchette, I.** (2018). Evidence for the role of cognitive resources in flavour–flavour evaluative conditioning. *Quarterly Journal of Experimental Psychology*, *65*(12), 2297–2308. DOI: https://doi.org/10.1080/17470218.2012.701311

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342–353. DOI: https://doi.org/10.1111/spc3.12111

Dedonder, J., Corneille, O., Bertinchamps, D., & **Yzerbyt, V.** (2014). Overcoming correlational pitfalls: Experimental evidence suggests that evaluative conditioning occurs for explicit but not implicit encoding of CS–US pairings. *Social Psychological and Personality Science*, *5*(2), 250–257. DOI: https://doi.org/10.1177/1948550613490969

Dedonder, J., Corneille, O., Yzerbyt, V., & **Kuppens, T.** (2010). Evaluative conditioning of high-novelty stimuli does not seem to be based on an automatic form of associative learning. *Journal of Experimental Social Psychology*, *46*(6), 1118–1121. DOI: https://doi.org/10.1016/j.jesp.2010.06.004

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(e33400), 1507–1517. DOI: https://doi.org/10.3389/fpsyg.2014.00781

Eagly, A. H., & **Chaiken, S.** (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Flake, J. K., & **Fried, E. I.** (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. DOI: https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & **Hehman, E.** (2017). Construct validation in social and personality research. *Social Psychological and Personality Science*, *8*(4), 370–378. DOI: https://doi.org/10.1177/1948550617693063

Gawronski, B., & **Bodenhausen, G. V.** (2014). Implicit and explicit evaluation: A brief review of the associative–propositional evaluation model. *Social and Personality Psychology Compass*, *8*(8), 448–462. DOI: https://doi.org/10.1111/spc3.12124

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. DOI: https://doi.org/10.1016/j.jesp.2012.01.002

Greenwald, A. G., & **Banaji, M. R.** (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. DOI: https://doi.org/10.1037//0033-295X.102.1.4

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139. DOI: https://doi.org/10.2307/2086306

Hofmann, W., Houwer, J. D., Perugini, M., Baeyens, F., & **Crombez, G.** (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. DOI: https://doi.org/10.1037/a0018916

Högden, F., Hütter, M., & **Unkelbach, C.** (2018). Does evaluative conditioning depend on awareness? Evidence from a continuous flash suppression paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(10), 1641–1657. DOI: https://doi.org/10.1037/xlm0000533

Hussey, I., & **Hughes, S.** (2019). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. DOI: https://doi.org/10.1177/2515245919882903

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. DOI: https://doi.org/10.1037/a0014747

Jurchis, R., Costea, A., Dienes, Z., Miclea, M., & **Opre, A.** (2020). Evaluative conditioning of artificial grammars: Evidence that subjectively unconscious structures bias affective evaluations of novel stimuli. *Journal of Experimental Psychology: General*, *149*(9), 1800–1809. DOI: https://doi.org/10.1037/xge0000734

Kattner, F. (2012). Revisiting the relation between contingency awareness and attention: Evaluative conditioning relies on a contingency focus. *Cognition & Emotion*, *26*(1), 166–175. DOI: https://doi.org/10.1080/02699931.2011.565036

Knowlton, B. J., Ramus, S. J., & **Squire, L. R.** (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science*, *3*(3), 172–179. DOI: https://doi.org/10.1111/j.1467-9280.1992.tb00021.x

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. DOI: https://doi.org/10.1177/2515245918771304

Levey, A. B., & **Martin, I.** (1975). Classical conditioning of human "evaluative" responses. *Behaviour Research and Therapy*, *13*(4), 221–226. DOI: https://doi.org/10.1016/0005-7967(75)90026-1

Lovibond, P. F., & **Shanks, D. R.** (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26. DOI: https://doi.org/10.1037//0097-7403.28.1.3

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314. DOI: https://doi.org/10.1177/014662169401800402

Mitchell, C. J., De Houwer, J., & **Lovibond, P. F.** (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183–198. DOI: https://doi.org/10.1017/S0140525X09000855

Moors, A., & **De Houwer, J.** (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. DOI: https://doi.org/10.1037/0033-2909.132.2.297

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K., Balas, R., Benedict, T., Corneille, O., Douglas, S. B., Ferguson, M. J., Fritzlen, K. A., Gast, A., Gawronski, B., Giménez-Fernández, T., Hanusz, K., Heycke, T., Högden, F., ... De Houwer, J. (2021). Incidental attitude formation via the surveillance task: A preregistered replication of the Olson and Fazio (2001) study. *Psychological Science*, *32*(1), 120–131. DOI: https://doi.org/10.1177/0956797620968526

Morey, R. D., & **Rouder, J. N.** (2018). *BayesFactor: Computation of Bayes Factors for common designs*. R package version 0.9.12–4.2. https://CRAN.R-project.org/package=BayesFactor

Newell, B. R., & **Shanks, D. R.** (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *38*(01), 1–19. DOI: https://doi.org/10.1017/S0140525X12003214

Olson, M. A., & **Fazio, R. H.** (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, *12*(5), 413–417. DOI: https://doi.org/10.1111/1467-9280.00376

Rouder, J. N., & **Morey, R. D.** (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689. DOI: https://doi.org/10.3758/s13423-011-0088-7

Rydell, R. J., McConnell, A. R., Mackie, D. M., & **Strain, L. M.** (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958. DOI: https://doi.org/10.1111/j.1467-9280.2006.01811.x

Saffran, J. R., & **Kirkham, N. Z.** (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. DOI: https://doi.org/10.1146/annurev-psych-122216-011805

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & **Botvinick, M. M.** (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. DOI: https://doi.org/10.1038/nn.3331

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*(3), 752–775. DOI: https://doi.org/10.3758/s13423-016-1170-y

Shanks, D. R., & **St. John, M. F.** (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*(03), 367–395. DOI: https://doi.org/10.1017/S0140525X00035032

Stahl, C., Haaf, J., & **Corneille, O.** (2016). Subliminal evaluative conditioning? Above-chance CS identification may be necessary and insufficient for attitude learning. *Journal of Experimental Psychology: General*, *145*(9), 1107–1131. DOI: https://doi.org/10.1037/xge0000191

Sweldens, S., Corneille, O., & **Yzerbyt, V.** (2014). The role of awareness in attitude formation through evaluative conditioning. *Personality and Social Psychology Review*, *18*(2), 187–209. DOI: https://doi.org/10.1177/1088868314527832

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & **Shanks, D. R.** (2019). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, *149*(1), 160–181. DOI: https://doi.org/10.1037/xge0000632

Valentine, J. C., Pigott, T. D., & **Rothstein, H. R.** (2009). How many studies do you need? *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. DOI: https://doi.org/10.3102/1076998609346961

Vazire, S. (2019). *"Thoughts inspired by the @replicats workshop: Replicability of Evidence asks 'Would I get consistent evidence if I did the same thing again?' Replicability of Inferences asks 'Would others draw the same inference from this evidence as the claim in the paper?' (1/5)."* [Tweet]. Twitter. https://twitter.com/siminevazire/status/1148149981292978178

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. DOI: https://doi.org/10.18637/jss.v036.i03

Waroquier, L., Abadie, M., & **Dienes, Z.** (2020). Distinguishing the role of conscious and unconscious knowledge in evaluative conditioning. *Cognition*, *205*, 104460. DOI: https://doi.org/10.1016/j.cognition.2020.104460

Wickham, H., François, R., Henry, L., & **Müller, K.** (2020). *dplyr: A grammar of data manipulation*. R package version 0.8.4. https://CRAN.R-project.org/package=dplyr

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. DOI: https://doi.org/10.1017/S0140525X20001685

TO CITE THIS ARTICLE:
Kurdi, B., Hussey, I., Stahl, C., Hughes, S., Unkelbach, C., Ferguson, M. J., & Corneille, O. (2022). Unaware Attitude Formation in the Surveillance Task? Revisiting the Findings of Moran et al. (2021). *International Review of Social Psychology*, 35(1): 6, 1–16. DOI: https://doi.org/10.5334/irsp.546

Submitted: 11 December 2020 **Accepted:** 03 May 2022 **Published:** 06 June 2022