“Evaluative Conditioning without awareness” is likely an artifact of failing to exclude aware participants:

A commentary on Moran et al. (2020) ‘Incidental Attitude Formation via the Surveillance Task: A Pre-Registered Replication of Olson and Fazio (2001)’

Ian Hussey & Sean Hughes

*Ghent University*

Abstract…..

The key verbal hypothesis being tested by the RRR is whether participants can demonstrate a change in liking due to the pairing of stimuli (i.e., Evaluative Conditioning, REF), when those pairing occur without awareness. Given that Evaluative Conditioning is uncontroversially known to occur within awareness (Hofmann et al., REF), testing this hypothesis requires us to successfully measure awareness and exclude or control for participants who are aware. In our opinion[[1]](#footnote-1), the demonstration of a significant effect in the original RRR’s confirmatory analysis was due the failure or the exclusion criterion to exclude individuals who are in fact aware, with the observed Evaluative Conditioning effect therefore being driven by these participants.

## Measuring awareness

The RRR considered differences in the Evaluative Conditioning effect produced by four different awareness exclusion criteria (“Olson & Fazio, 2001”, “Olson & Fazio, 2001 modified”, “Bar-Anan, De Houwer & Nosek, 2010”, and “Bar-Anan, De Houwer & Nosek, 2010 modified”; see Moran et al., 2020, for details). Results from the first criteria, which replicate those used in the original study by Olson & Fazio (2001), were used as the confirmatory analysis in order to determine whether the effect successfully replicated or not. We note that a significant effect was found using this criterion (Hedges’ *g* = 0.12, 95% CI [0.05, 0.20], *p* = .002) but not the other criteria (see Moran et al., 2020). However, Moran et al. (2020) did not consider the validity of the awareness exclusion criteria. Here, we provide evidence that suggests that the awareness criteria – and in particularly the Olson & Fazio, 2001 criterion used for confirmatory analyses – are noisy measures of awareness with poor validity, which may lead to incorrect inferences.

It should be noted that the Olson & Fazio (2001) criterion used for confirmatory analyses, and the only criterion under which a significant effect was found, was also the laxest criterion: it excluded only 8% of participants where others excluded up to 48% (Olson & Fazio, 2001 modified = 31%; Bar-Anan, De Houwer & Nosek, 2010 = 48%; Bar-Anan et al., 2010 modified = 27%). As well as this variability in exclusion rates between criteria, a great degree of variability in exclusion rates was also observed between data collection sites. For example, exclusion rates using the “Olson & Fazio (2001) modified” criterion varied between 15% and 74% between sites. In order to assess whether this variation could be attributed to mere sampling variation or, more problematically, due to between-site heterogeneity, we conducted meta analyses of the proportion of aware participants between sites for each criterion (see Supplementary Materials for full details and code). Results suggested that the variation in awareness rates between sites represented a large degree of between-site heterogeneity rather than merely sampling variation (across exclusion criteria all *I*2 = 54.7% to 91.7%, all *H*2 = 2.2 to 12). Such heterogeneity between sites may be attributable to genuine differences in participant samples between sites. However, more problematically, it could also be due to differences in scoring methods. This seems plausible given that, despite standardized instructions being provided to each site, scores were calculated from open-ended responses that were hand-scored by researchers at each site, making them less than objective. In our opinion, this heterogeneity makes it highly plausible that (1) at some sites, participants who were aware were not appropriately excluded, and (2) this lead to the observed Evaluative Conditioning effect, which was driven by aware participants. In the following sections, we attempted to control for this possibility in two different ways (statistically vs. methodologically), in order to provide more severe tests of the hypothesis.

## Statistically controlling for awareness

In order to statistically control for inter-site differences in the scoring of awareness, we conducted a meta-analysis of Evaluative Conditioning effects. This was highly similar to the confirmatory meta-analysis in Moran et al. (2020) but with two modifications. First, we made no exclusions based on awareness but instead used the full sample (*N* = 1450). Second, we controlled for the awareness rates at each site, using the Olsen & Fazio (2001) criterion (i.e., entered site awareness rate as a moderator). Results demonstrated that, when the awareness rate is statistically set to 0% (i.e., the model intercept), the estimate of the Evaluative Conditioning effect was non-significant and close to zero, Hedges’ *g* = -0.02, 95% CI [-0.35, 0.31], *p* = .223, and with no heterogeneity, 𝜏2 = 0.0, *I*2 = 0.0%, *H*2 = 1.0. These results support the idea that (1) the observed heterogeneity in awareness rates between sites may be due to the somewhat subjective nature of the awareness scoring criteria that may have differed between labs, and therefore (2) the presence of significant meta effect size in the original confirmatory analysis may have been due to the failure of this criteria to strictly exclude aware participants.

## Stricter exclusion of awareness

In order to methodologically control for awareness more strictly, we adopted a compound awareness exclusion criterion that represented a more severe test of the hypothesis: participants were excluded if *any* of the four criteria scored them as aware. Given that a compound exclusion strategy would increase exclusion rates beyond what was originally planned, we first conducted power analyses to assess whether a new analysis would provide meaningful results. The compound exclusion criterion produced an exclusion rate of 54%, leaving 665 participants in the analytic sample. Using the power analysis method employed in Moran et al. (2020), this sample size maintained power >.99 to detect an effect size as large as that observed in the published literature (i.e., *g* = 0.20, see Moran et al., 2020). It also had power = .80 to detect an effect size of as small as 0.10. We also considered an additional form of power analysis that acknowledged the hierarchical structure among the data given the multi-site design. This still demonstrated power = .95 to detect the effect size found in the literature, or to detect an effect size of as small as 0.16 with power = .80 (see Supplementary Materials for details and code). Using this compound exclusion criterion, a second meta-analysis demonstrated 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, and with no heterogeneity, 𝜏2 = 0.0, *I*2 = 0.0%, *H*2 = 1.0.

## Conclusions

While the RRR replicated the original *effect* demonstrated in Olson & Fazio (2001), we argue that both Olson & Fazio (2001) and the RRR represent weak tests of the underlying verbal hypothesis that Evaluative Conditioning can be demonstrated in the absence of awareness of stimulus pairings. Let us be clear: we are not arguing the effect produced by this experimental setup doesn’t replicate. The results of the RRR indicate that it does. Rather, we are arguing this experimental setup is not particularly useful for testing the verbal hypothesis of interest. In our opinion, this setup produces replicable effects but unreplicable inferences.

In our opinion, this is primarily due to the fact that the exclusion criteria appear to function as poor or noisy measures of awareness. Two new meta analyses, which (a) controlled for the awareness rate between sites and (b) used stricter compound awareness criteria, both demonstrated non-significant Evaluative Conditioning effects that were close to zero. In our opinion, when subjected to severe testing, the key hypothesis tested by the RRR is not supported by the data. Results serve to highlight the importance of distinguishing between a replicable effect and a replicable inference, as well as highlighting the need within experimental psychology to pay greater attention to measurement validity if our effects and inferences are to be both replicable and valid (see Flake REF; Hussey & Hughes, 2020).

## Author contributions

IH conceptualized the study and analyzed the data. SH provided critical input to the design and analysis. Both authors wrote the article and approved the final submitted version of the manuscript.

## Declaration of Conflicting Interests

IH and SH declare we have no conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted with the support of Ghent University grant 01P05517 to IH and BOF16/MET\_V/002 to Jan De Houwer.

# References

1. It is perhaps unconventional to write a commentary on a paper which we are third and second authors of (respectively), and therefore it’s worth providing a brief explanation. This situation arose in part due to the large number of authors involved in the original RRR paper who, understandably, possess a diverse set of opinions on how its results should be interpreted. Moran et al. (2020) represents the consensus opinion among that study’s authors, whereas this commentary provides our own opinions on the results of that study. [↑](#footnote-ref-1)