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

Ian Hussey & Sean Hughes

*Ghent University*

Moran et al.’s (2019) Registered Replication Report of Olson and Fazio’s (2001) study set out to test a seemingly simple verbal hypothesis: can changes in evaluative responding due to the pairing of stimuli (i.e., Evaluative Conditioning effects: ‘EC’) take place even when people are ‘unaware’ that stimuli have been paired? Moran et al. reported a significant effect in their confirmatory analysis which some might be tempted to see as support for the idea of ‘unaware EC’. We would strongly caution against such an interpretation. Testing the ‘unaware EC’ hypothesis requires a valid and reliable measure of awareness capable of excluding or controlling for ‘aware’ participants. In our opinion, the confirmatory effect obtained in Moran et al. (2019) was primarily driven by the fact that the exclusion criterion used in those analyses failed to exclude individuals who were aware, with the observed effect driven by these ‘aware’ participants. [[1]](#footnote-2)

To briefly recap, Moran et al. (2019) examined if EC effects on the surveillance task differed when four different awareness exclusion criteria were applied (i.e., the “Olson & Fazio, 2001”, “Olson & Fazio, 2001 modified”, “Bar-Anan, De Houwer & Nosek, 2010”, and “Bar-Anan, De Houwer & Nosek, 2010 modified” criteria; for details of each see Moran et al., 2019). Their confirmatory analysis was based on the original authors exclusion criterion (i.e., “Olson & Fazio, 2001”) which, when applied, led to a significant effect (Hedges’ *g* = 0.12, 95% CI [0.05, 0.20], *p* = .002). Applying any of the other three (exploratory, yet pre-registered) exclusion criteria did not lead to an EC effect.

What Olson and Fazio (2001; and by extension; Moran et al., 2019) failed to do, in our opinion, was to consider the *validity* of these four awareness exclusion criteria. Our commentary does so, and provides evidence that the four awareness criteria are noisy measures of awareness with poor validity, particularly the original authors criterion central to the confirmatory analyses. These noisy measures may lead to incorrect inferences about the underlying mechanisms being studied.

## Not all measures of are created equal

As we previously mentioned, the original authors’ (“Olson & Fazio, 2001”) criterion was used in the confirmatory analyses, and was the only criterion under which a significant effect was found. Yet it was also the most liberal criterion by far: it flagged a mere 8% of participants as being aware, whereas the other exclusion criteria signaled that up to 48% of participants were aware of the underlying manipulation (“Olson & Fazio, 2001 modified” = 31%; “Bar-Anan, De Houwer & Nosek, 2010” = 48%; “Bar-Anan et al., 2010 modified” = 27%). Not only was there great variation between the four exclusion criteria, but also large variability in exclusion rates within each criteria between data collection sites. For example, exclusion rates using the “Olson & Fazio (2001) modified” criterion varied between 15% and 74% between the data sites.

This led us to investigate if such variation was due to mere sampling variation or, more problematically, to between-site heterogeneity. We therefore conducted four meta analyses of the proportion of aware participants between sites, one for each of the exclusion criterion (see Supplementary Materials for full details, code, and results of all analyses). Results indicated that variation in awareness rates between sites was due to a large degree of between-site heterogeneity rather than mere sampling variation (across exclusion criteria all *I*2 = 54.7% to 91.7%, all *H*2 = 2.2 to 12). What exactly does this mean? Well, such heterogeneity between sites may be due to genuine differences in participant samples between sites (i.e., how aware people were of the EC manipulation in different labs). Yet it could also be due to differences in scoring methods across exclusion criteria. 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 this process highly subjective.

In our opinion, the combination of (1) clear differences in the strictness of the criteria and (2) heterogeneity within each criterion between sites makes it highly plausible that participants who were aware were not appropriately excluded, and that this led to the observed EC effect in the confirmatory condition. If we want to conclude that EC effects can emerge in the absence of awareness, we require a more severe test of the verbal hypothesis.

## Statistically controlling for awareness

To statistically control for awareness than in Olson & Fazio (2001) and Moran et al. (2019), we conducted a moderator meta-analysis of EC effects. This was highly similar to the confirmatory meta-analysis in Moran et al. (2019), 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 EC 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 subjective nature of the awareness scoring criteria which may have differed across 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

To better methodologically control for awareness, we combined the four exclusion criterion to create a stricter, compound exclusion criterion: participants were excluded if *any* of the four criteria scored them as aware. By excluding participants that were scored as aware on any criterion, this provided a more severe test of the hypothesis. Given that a compound exclusion criterion would increase exclusion rates beyond what was originally planned, we first conducted power analyses. The compound exclusion criterion flagged 54% of participants as aware to some degree and thus left 665 participants in the analytic sample. Using the power analysis method employed by Moran et al., 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). It also had power = .80 to detect an effect size > 0.10. In addition, we also considered a further type of power analysis that acknowledged the hierarchical structure among the data given the multi-site design. These also suggested we had power = .95 to detect the effect size commonly reported in the literature, or an effect size as small as 0.16 with power = .80 (see Supplementary Materials for details and code). After the compound exclusion criterion was applied, 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, and with no heterogeneity, 𝜏2 = 0.0, *I*2 = 0.0%, *H*2 = 1.0.

## Conclusions

While Moran et al. (2019) replicated the original *effect* reported in Olson & Fazio (2001), we argue that both Olson & Fazio (2001) and Moran et al. (2019) represent weak tests of the underlying verbal hypothesis of unaware EC. Let us be clear: we are not arguing the EC effect produced by the surveillance task does not replicate. The results of the RRR indicate that it does. Rather, we are arguing that such an experimental setup is a poor test of the verbal hypothesis that is ultimately of interest (i.e., EC effects in the absence of ‘awareness’). In our opinion, the surveillance task and awareness measures used produce replicable effects but unreplicable inferences.

This is mainly 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 EC effects that were close to zero. In our opinion, when subjected to more rigorous testing, the key hypothesis tested by Moran et al. (2019) 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 to pay greater attention to measurement validity if our effects and inferences are to be both replicable and valid. Such calls have been made within other areas of psychology (see Flake et al., 2017; Hussey & Hughes, 2018), but rarely within experimental psychology.

## Author contributions

IH conceptualized the study and analyzed the data. SH provided critical input into 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

Bar-Anan, Y., Houwer, J. D., & 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. https://doi.org/10.1080/17470211003802442

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Hussey, I., & Hughes, S. (2018). *Hidden invalidity among fifteen commonly used measures in social and personality psychology*. https://doi.org/10.31234/osf.io/7rbfp

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K., Balas, R., Benedick, T., Corneille, O., Douglas, S. B., Ferguson, M. J., Fritzlen, K. A., Gast, A., Gawronski, B., Heycke, T., Högden, F., Hütter, M., Kurdi, B., … De Houwer, J. (2019). Incidental Attitude Formation via the Surveillance Task: A Pre-Registered Replication of Olson and Fazio (2001). *Psychological Science*, *Registered Replication Report Stage 1 acceptance*.

Olson, M. A., & Fazio, R. H. (2001). Implicit Attitude Formation Through Classical Conditioning. *Psychological Science*, *12*(5), 413–417. https://doi.org/10.1111/1467-9280.00376

1. It may seem unconventional to write a commentary on a paper which we are third and second authors of (respectively). This situation arose in part due to the large number of authors involved in Moran et al. (2019) who, understandably, possess a diverse set of opinions on the concept of ‘awareness’ and how results in that article 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-2)