

Evaluative Conditioning without awareness: Replicable effects do not equate replicable inferences

Ian Hussey & Sean Hughes
Ghent University

Moran et al. (2019) successfully replicated the surveillance task effect obtained by Olson & Fazio (2001). This effect is often treated as evidence for attitude formation in the absence of awareness. However, such an inference requires that ‘aware’ participants are successfully excluded from consideration. We present evidence that the awareness exclusion criterion used by Olson & Fazio (2001) – and the only one to produce a significant effect in Moran et al. (2019) – is a poor measure of awareness: it is overly lax, noisy, and demonstrates heterogeneity between sites. Two new meta-analyses, which (a) used stricter compound awareness criteria and (b) controlled for differences in awareness rates between sites, both demonstrated non-significant, near-zero effect sizes. When subjected to more severe testing, Moran et al.’s (2019) data does not support the ‘unaware Evaluative Conditioning’ hypothesis.

Olson and Fazio (2001) presented evidence that changes in evaluative responding due to the pairing of stimuli (i.e., Evaluative Conditioning effects: ‘EC’) can take place even when people are ‘unaware’ that stimuli have been paired. Recently, Moran et al. (2019) conducted a close replication.¹ While Moran et al.’s (2019) results 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 Olson and Fazio’s (2001) surveillance task does not replicate. The results of Moran et al.’s (2019) 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 produced replicable *effects*, but unreplicable *inferences* regarding the verbal

hypothesis of ‘unaware Evaluative Conditioning’ (see Hussey & Hughes, 2020; Yarkoni, 2019).

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 et al. (2010)”, and “Bar-Anan et al. (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 a significant EC effect.

Of course, testing the ‘unaware EC’ hypothesis requires a valid and reliable measure of awareness capable of excluding participants who were aware of the stimulus pairings. What Olson and Fazio (2001; and by extension Moran et al., 2019) failed to do, in our opinion, was to

¹ We are third and second authors (respectively) of Moran et al. (2019). Given the large number of authors involved in Moran et al. (2019), there was a diverse set of opinions on the concept of

‘awareness’ and how results in that article should be interpreted. Moran et al. (2019) represents the consensus opinion among that study’s authors, whereas this commentary provides our own opinions.

consider the *validity* of these four awareness exclusion criteria. In our opinion, the confirmatory effect obtained in Moran et al. (2019) was primarily driven by the fact that the exclusion criterion used in that analysis failed to exclude individuals who were aware, with the observed effect driven by these aware participants. Here we provide (1) evidence that the four awareness criteria are noisy measures of awareness with poor validity, and (2) stricter tests of the core hypothesis.

Not all measures of awareness 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. What Moran et al. (2019) fails to emphasize is that this criterion was also the most liberal criterion by far: it scored only 8% of participants as aware, whereas other exclusion criteria scored up to 48% of participants as aware ("Olson & Fazio, 2001 modified" criterion = 31%; "Bar-Anan et al., 2010" criterion = 48%; "Bar-Anan et al., 2010 modified" criterion = 27%).

It is important to note that these criteria were not merely different in their strictness, but also produced incongruent decisions between criteria. For example, awareness scorings between the "Olson & Fazio (2001) modified" and "Bar-Anan et al. (2010) modified" criteria were incongruent in 20.8% of participants. Across all four criteria, violations of a unidimensional structure were common (Guttman errors: $G = 47.5\%$, 95% CI [45.5, 49.5], $G^* = 11.9\%$, 95% CI [11.4, 12.4]; Meijer, 1994. See Supplementary Materials for full details, data, code, and results of all analyses).

Not only was there great variation between the four exclusion criteria, but also large differences 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 sites. Meta-analyses of the proportion of aware participants between sites for each of the exclusion criteria demonstrated large between-site heterogeneity (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, more problematically, between-site heterogeneity. Given that all measures and instructions were provided to each site in a standardized format, this heterogeneity in results provides further evidence that the awareness measures may not be as valid as previously assumed.² This could be attributed, at least in part, to the somewhat subject

nature of the "Olson & Fazio (2001)" criterion, which involves hand scoring participants' open-ended responses.

The combination of (1) clear differences in the strictness of the criteria combined with the confirmatory analysis's reliance on the most lax criterion, (2) incongruence between the conclusions reached between criteria, and (3) heterogeneity within each criterion between sites, makes it highly plausible that the awareness measures demonstrate poor validity and likely failed to exclude participants who were aware; and it is this that this led to the significant effect in Moran et al.'s (2019) confirmatory condition. If we want to conclude that EC effects be demonstrated in the absence of awareness, we require more severe tests of the verbal hypothesis.

Stricter exclusion of awareness

To more strictly exclude aware participants, we simply combined the four exclusion criterion to create a compound exclusion criterion. That is, participants were excluded if *any* of the four criteria scored them as aware. This provided a more severe test of the hypothesis. This compound exclusion criterion scored 54% of participants as aware to some degree, leaving 665 participants in the analytic sample. Using the power analysis method employed by Moran et al. (2019), this sample size still 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 . After applying the stricter compound exclusion criterion, 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], $p = .983$.

Statistically controlling for awareness

Given that the core hypothesis concerns EC effect in the absence of awareness, and there was heterogeneity in awareness rates between sites, we conducted a moderator meta-analysis of EC effects that controlled for awareness rate. 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 "Olson & Fazio (2001)" criterion (i.e., entered site awareness rate as a moderator). Results demonstrated that the 'unaware EC' effect (i.e., when site awareness was 0%) was non-significant and close to zero, Hedges' $g = -0.02$, 95% CI [-0.35, 0.31], $p = .223$.

and as such we are highly familiar them and the efforts to standardize them between sites.

² It is worth noting that the first author was responsible for the creation and distribution of the measures used in Moran et al. (2019),

Conclusions

Olson & Fazio's (2001) study and Moran et al.'s (2019) replication of their effect both rely on the successful exclusion of aware participants. However, neither study assessed the validity of their awareness criteria. Our analyses suggest that the criteria are, individually, relatively poor measures of awareness that likely fail to exclude aware participants. When subjected to more severe testing, Moran et al.'s (2019) data does not support the 'unaware Evaluative Conditioning' hypothesis. Results serve to highlight the importance of distinguishing between a replicable *effect* and a replicable *inference* regarding the verbal hypothesis, as well as highlighting the need to pay greater attention to measurement validity if our 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, 2020), but rarely within experimental psychology.

Notes

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

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. (2020). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *In Press*. <https://doi.org/10.31234/osf.io/7rbfp>
- Meijer, R. R. (1994). The Number of Guttman Errors as a Simple and Powerful Person-Fit Statistic. *Applied Psychological Measurement*, 18(4), 311–314. <https://doi.org/10.1177/014662169401800402>
- 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*. <https://osf.io/hs32y>
- 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>
- Yarkoni, T. (2019). *The Generalizability Crisis. Preprint*. <https://doi.org/10.31234/osf.io/jqw35>