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

Ian Hussey & Sean Hughes

Moran et al.'s (2020) primary analysis successfully replicated the surveillance task effect obtained by Olson and 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 and Fazio (2001) – the only one to produce a significant effect in the replication – is a poor measure of awareness: it is overly lax, noisy, and demonstrates heterogeneity between sites. A new meta-analysis using a stricter compound awareness criterion that prioritized sensitivity (N = 665) demonstrated a non-significant and near-zero effect size (Hedges' g = 0.00, p = .983, BF_{10} = 0.04). When subjected to a more severe test, Moran et al.'s (2020) data does not support the 'unaware Evaluative Conditioning' hypothesis. Results serve to highlight the importance of distinguishing between a replicable statistical *effect* and a replicable *inference* regarding a verbal hypothesis.

Olson and Fazio (2001) presented evidence that changes in liking 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. (2020) conducted a close replication of this work.[1] While Moran et al.'s (2020) results replicated the original effect reported in Olson and Fazio (2001), we argue that both Olson and Fazio (2001) and Moran et al. (2020) 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. (2020) indicate that it does. Rather, we are arguing that this experimental setup is a poor test of the verbal hypothesis that is ultimately of interest. In our opinion, the surveillance task and awareness measures produced replicable statistical *effects,* but unreplicable *inferences* regarding the verbal hypothesis of 'unaware Evaluative Conditioning' (distinction originally made by Vazire, 2019; see also Hussey & Hughes, 2020; Yarkoni, 2019).

To briefly recap, Moran et al. (2020) examined if EC effects on the surveillance task were present when four different awareness[2] exclusion criteria were applied (i.e., the 'Olson & Fazio, 2001', 'Olson & Fazio, 2001 modified', 'Bar-Anan, De Houwer, & Nosek, 2010', and 'Bar-Anan et al., 2010 modified' criteria; for details of each see Moran et al., 2020). Their primary 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 (pre-registered) secondary exclusion criteria did not lead to significant EC effects (all gs = 0.03 to 0.05, all ps > .241).

Of course, testing the 'unaware EC' hypothesis requires a reliable and valid measure capable of excluding participants who were 'aware' of the stimulus pairings. What Olson and Fazio (2001) failed to do, in our opinion, was to consider the structural validity of these four awareness exclusion criteria. While Moran et al. (2020) noted that "any attempt to detect differences

[1] We are third and second authors (respectively) of Moran et al. (2020). Given the large number of authors involved in Moran et al. (2020), there was 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.

[2] As Moran et al. (2020) note, there is debate as to whether the exclusion criteria capture 'awareness' of the stimulus pairings, 'recollective memory' of this awareness, or both (see Gawronski & Walther, 2012; Jones et al., 2009). Here we refer to the criteria as measures of awareness throughout the current article. Rather than focus on what is being measured, we focus on the more fundamental question of whether they are reliable measures in the first place.

in EC effects between putatively 'aware' and 'unaware' participants will ultimately depend on the reliability of the awareness measure" (p. 23), and that such measures are frequently unreliable (Shanks, 2017; Vadillo et al., 2020), that article did not contain any direct consideration of the structural validity of the awareness measures. Recent work has argued that such issues around measurement are common yet underappreciated in psychology and serve to threaten the validity of our findings and the conclusions we draw from them (Flake et al., 2017; Flake & Fried, 2019; Hussey & Hughes, 2020).

In our opinion, the effect obtained in Moran et al.'s (2020) primary analysis was 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. In this paper we (1) assess the validity of the four awareness criteria and conclude that they are poor and noisy measures of awareness, and (2) conduct a stricter test of the core verbal hypothesis and conclude that the evidence does not support 'unaware EC'.

Poor measures of awareness

Reliability between criteria

As we previously mentioned, the 'Olson and Fazio (2001)' criterion used in the primary analysis was the only criterion under which a significant EC effect was found. Importantly, it was also the most liberal one 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%). While these awareness rates were reported in Moran et al. (2020), that article did not directly consider the relationship between the criteria's relative strictness and the EC effects they produced.

What the above shows is that there were meaningful differences in the exclusion rates observed between criteria. In an everyday sense, these measure would differ only in their relative strictness, rather than there also being unreliability between them. More formally, 'strictness' in this context is a quantifiable statistical property referred to as the degree of conformity to a Guttman structure, which is testable using methods from Item Response Theory modelling. Specifically, if these measures demonstrated perfect reliability and differed only in strictness we would expect the proportion of Guttman errors (G) to be very small (i.e., approach 0). In contrast, if they were unreliable we would expect G to approach 1.

Results demonstrated that measures were indeed quite unreliable. Nearly half of participants had scores on one or more awareness criteria that indicated such errors, G = 47.5%, 95% CI [45.5, 49.5], G^* = 11.9%, 95% CI [11.4, 12.4] (see Meijer, 1994; and see Supplementary Materials for full details, data, code, and results of all analyses). In other words, in about half of participants, a supposedly more lenient awareness criterion actually scored them more strictly than a supposedly stricter criterion.

Reliability between sites

There was also a great deal of variation in the exclusion rates between data collection sites. For example, exclusion rates using the 'Olson and Fazio (2001) modified' criterion varied between 15% and 74% between sites. This was quantified using meta-analyses of the proportion of 'aware' participants between sites for each of the exclusion criteria. Results demonstrated large between-site heterogeneity (all I^2 = 54.7% to 91.7%, all H^2 = 2.2 to 12 between the four criteria). Differences in between-site awareness rates therefore did not represent mere sampling variation but rather large between-site heterogeneity. Given that all measures and instructions were delivered to participants in a standardized format, this degree of heterogeneity represents evidence that the awareness measures may not be as reliable or valid as assumed.[3]

This 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) require these to be hand scored. This method leaves room for a great degree of variation in interpretation between participants and sites which ultimately could lead many 'aware' participants to be scored as 'unaware'. To take just one example, an individual who is fully 'aware' of the pairings in the surveillance task might reasonably consider the stimulus pairings to be unremarkable and not odd at all, but merely a normal and obvious part of the task, respond as such, and therefore be scored incorrectly as 'unaware'.

The preceding two sections suggest that the awareness criteria demonstrated poor reliability and structural validity, and therefore likely failed to exclude participants who were actually aware. In our opinion, it was this that this led to the significant effect in Moran et al.'s (2020) primary analysis (i.e., its reliance on the worst of a bad bunch). If we want to conclude that EC effects can be demonstrated in the absence of awareness, then a more severe test of the verbal hypothesis is required.

[3] It is worth noting that the first author was responsible for the creation and distribution of the measures used in Moran et al. , and as such is highly familiar with them and the efforts to standardize them between sites.

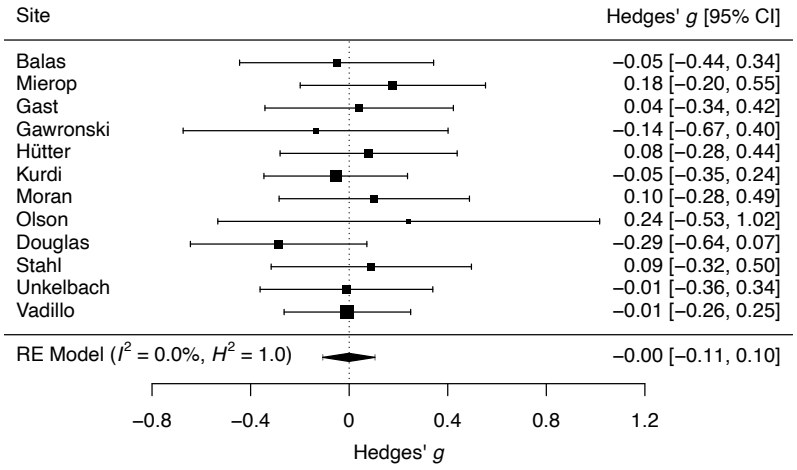

Figure 1. Forest plot of results.

A severe test of the 'unaware EC' hypothesis

With this in mind, we created a stricter exclusion criterion that favored sensitivity over specificity, and therefore maximized our chances of excluding 'aware' participants. Specifically, we excluded participants if *any* of the four criteria scored them as being aware. This compound criterion excluded 54% of participants as 'aware', leaving 665 in the analytic sample.

Before fitting a new meta-analysis model, we first assessed the statistical power of this test given the available sample size. This ensured that the results of such a test would be meaningful. Using the same power analysis method employed by Moran et al. (2020), to detect an effect size as large as that observed in the published literature (i.e., $g = 0.20$) with this sample size, power was $> .99$. Stated another way, at power = .80, the minimum detectable effect size was Cohen's $d = 0.10$. Power estimates were comparable when we employed what we considered to be a more appropriate method of power analysis for meta-analysis models (see Valentine et al., 2010): to detect an effect size of $d = 0.20$, power was = .95. At power = .80, the minimum detectable effect size was $d = 0.16$. The available sample size was therefore concluded to demonstrate adequate statistical power for our analysis, comparable to Moran et al. (2020).

After excluding participants using the compound criterion, we fitted a meta-analysis model that was otherwise identical to that employed in Moran et al.'s (2020) 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-analysis model using Rouder and Morey's (2011) method was also fitted 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 fixed effect for condition and $r = 1.0$ on random effect for data collection site, see Rouder & Morey 2011). Strong evidence was found in favor of the null hypothesis ($BF_{10} = 0.04$, effect size $\delta = 0.00$, 95% HDI [-0.08, 0.07]).

Conclusions

Olson and Fazio's (2001) study and Moran et al.'s (2020) replication both rely on the successful exclusion of 'aware' participants. However, neither study assessed the reliability or 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. We created a stricter awareness exclusion criterion that prioritized sensitivity by combining all four into a compound exclusion criterion. When subjected to this more severe test, Moran et al.'s (2020) data does not support the 'unaware Evaluative Conditioning' hypothesis.

Results serve to highlight the importance of distinguishing between a replicable statistical *effect* and a replicable *inference* regarding a verbal hypothesis of interest (Vazire, 2019; see Yarkoni, 2019), as well as highlighting the need to pay greater attention to measurement 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; Flake & Fried, 2019; Hussey & Hughes, 2020), but rarely within experimental social psychology.

Finally, as coauthors of Moran et al. (2020), we regret that we did not consider creating this compound criterion prior to the preregistration of the replication. Preregistration prior to seeing the results of the primary tests would have increased the evidential weight of the current results. However, the concept of evidential weight is at the core of our critique here: as Moran et al. (2020) note in their discussion, claims for the

replicability of support for the verbal hypothesis of 'unaware EC' have far reaching implications, and such claims require strong evidence. We feel that the general trend of evidence, across Moran et al.'s (2020) analyses and those reported here, is against 'unaware EC'.

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., & Fried, E. I. (2019). *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Preprint.* https://doi.org/10.31234/osf.io/hs7wm

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

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

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

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

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., Giménez-Fernández, T., Hanusz, K., Heycke, T., Högden, F., ... De Houwer, J. (2020). 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

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

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. https://doi.org/10.3758/s13423-016-1170-y

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (2020). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General*, *149*(1), 160–181. https://doi.org/10.1037/xge0000632

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. 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]. https://twitter.com/siminevazire/status/1148149981292978178

Yarkoni, T. (2019). *The Generalizability Crisis. Preprint.* https://doi.org/10.31234/osf.io/jqw35