Dear Prof. Lindsay,

Please find attached the revised version of MS number PSCI-19-0402, (RR-PDR) titled “Incidental Attitude Formation via the Surveillance Task: A Pre-Registered Replication of Olson and Fazio (2001)”. We thank you and Reviewers 1-2 for your positive and valuable feedback which we have used to further improve our paper. You can find our detailed responses to that feedback below.

Best Regards,

Coordinating Team:

Tal Moran (Tal.MoranYorovich@ugent.be)

Sean Hughes (sean.hughes@ugent.be)

Ian Hussey (ian.hussey@ugent.be)

Miguel A. Vadillo (miguel.vadillo@uam.es)

Jan De Houwer (Jan.DeHouwer@ugent.be)

**REVIEWER 1’S COMMENTS**

**Reviewer 1**: The goal of the pre-registered replication proposal is to test whether evaluative conditioning can occur without conscious awareness. Per instructions for reviewing a preregistered replication, I will explicitly address the four questions posed in the request for this review.

**[R1.1]** Is the original effect sufficiently important, interesting, impactful, and plausible yet uncertain to warrant direct replication? Why or why not?

I think the original effect is sufficiently important, interesting, etc. that it warrants direct replication. The original paper has been cited more than 600X demonstrating considerable interest, and the authors of the pre-registered replication proposal present a strong case for the importance of knowing just how robust this effect is for future theoretical efforts as well as practical application of these findings. The meta-analysis of existing studies was particularly helpful for bringing into focus that this finding could be the result of publication bias, so a large scale, multi-lab replication effort seems warranted.

**Authors**: We would like to thank Reviewer 1 for the positive evaluation of our work and recognition of its potential importance.

**Reviewer 1:** If it is, then is the proposed study the best way to conduct and analyze such a replication? Is the proposal sufficiently detailed and complete for you to answer this question?

Looping in the original authors of the evaluative conditioning effect as well as bringing together a diverse set of different labs seems to me to be a very strong way to conduct this replication attempt.

**[R1.2]** My major questions or point of pause were the caveat “whenever possible…” the labs would use the specifically targeted stimuli. Under what conditions precisely would labs deviate from the proposed method? In addition, would all studies be conducted in English and therefore the study participants screened for English language skills, or is the plan to translate the materials into the dominant language where the study is being conducted? If it is the latter, a lot more needs to be elaborated about the details of translation (and back translation) would be handled, and what steps might be taken to ensure the fidelity of the experimental process would be preserved (e.g. measurement equivalence).

**Authors**: Reviewer 1 had a number of comments which we address separately below:

*Comment #1 : Stimulus selection and deviations from the original authors work*. The “whenever possible” phrase that Reviewer 1 mentions refers to the selection process via which we choose the stimuli for our pre-registration (and not to the materials that the participating labs will ultimately use).

On the one hand, we were not able to use the original Pokémon characters (i.e., the conditioned stimuli [CSs], filler and target stimuli) that the original authors used because the original authors indicated, both during our correspondence and in their published work (Jones, Olson, & Fazio, 2009) that most of those characters are highly familiar to participants. This is problematic for two reasons. First, familiarity may increase the salience of a given character, draw attention towards it, and thus undermine the so-called ‘implicitness’ of the procedure. Second, any changes in liking towards those characters may not reflect evaluative conditioning *per se*, but rather pre-existing preferences for one character over another. In order to control for this, we pre-rated a set of recently developed Pokémon characters in terms of their valence and familiarity, and then selected the subset of characters that were both neutral and non-familiar to participants.

On the other hand, we were able to use nineteen of the twenty positive items and nineteen of the twenty negative items (i.e., unconditioned stimuli; USs) used in the original authors work. We had to replace one positive and one negative item because the quality of those images was simply too poor. We discussed this with the original authors and they approved the two replacement images used in our pre-registration.

Finally, with regard to the filler items (i.e., the neutral words and images), the original authors did not provide us with filler items and we therefore had to select these items ourselves. That said, we ensured that those items were approved by the original authors before they were included in our pre-registered study.

In short, we used the phrase ‘whenever possible’ in our original submission to capture the above points (i.e., deviations in selection of stimulus materials relative to the original authors’ work). We have now revised the materials section in our manuscript to better clarify and communicate these deviations (see changes on pp. 5-7).

*Comment #2: Stimulus selection and deviations across labs*. Reviewer 1 also asked about deviations in the exact stimuli used within and between participating laboratories. On the one hand, all laboratories will use the exact same set of unconditioned stimuli (USs) and neutral filler stimuli. On the other hand, labs will differ in the specific Pokémon characters (i.e., the CSs, filler, and target stimuli) to ensure that those characters are both neutral and unfamiliar to participants in each specific location. We now elaborate more on this point in the revised manuscript (see changes on p.6):

“We pretested these characters along two dimensions (valence and familiarity) with a separate sample of 155 participants using the Prolific Academic website (https://prolific.ac) (see https://osf.io/2f94r/). On the basis of this pretest we then selected the twenty characters that were rated as most neutral and least familiar. Participating labs will be instructed to further pretest these twenty characters onsite in order to identify the nine characters that are most neutral and least familiar to participants at that specific lab. The two characters that (a) are most neutral and least familiar, and (b) which differ least in valence and familiarity will serve as CSs. Labs unable to carry out such a pretest will use the nine characters derived from our own initial pretest. In this case two characters (Palpitoad and Bergmite) will serve as CSs.”

*Comment #3: Translation and fidelity process*. Reviewer 1 requested additional information on how materials will be translated within and between labs. We have now included such information in our revised paper (see changes on pp.10-11):

“**Experimental fidelity.** We have taken a number of steps in order to maximize experimental fidelity across labs. First, given differences in the native languages of participating labs (e.g., Dutch, German, Spanish, French, Polish), materials originally produced in English will be translated. We will do so using a forward and backward translation process. Specifically, materials will first be translated from English into the native language used at a given lab by one member of that participating team. This translation will then be backward translated into English by another member of that same team who was not involved in the initial translation process. This backward translation will be returned to the coordinating team for verification and approval. If necessary (i.e., where the backward translation is not approved) the translation process will be repeated until approval is provided.”

We have also taken further steps to maximize the fidelity of our experimental process. First, we have designed the study in such a way that the entire experimental protocol is standardized across all participating labs. Specifically, each lab will run the experiment using the same program and general materials (i.e., in PsychoPy, Peirce, [2007](https://www.tandfonline.com/doi/full/10.1080/02699930903485076)) which will generate identically formatted raw data files across all sites. We will then collate these data files from all sites and analyze them using the same R code and scripts. All materials and analytic files will be pre-registered before data collection begins. We now acknowledge this in the revised manuscript (see changes on p.11):

“... the entire experimental protocol will be standardized across all labs. Specifically, each lab will run the experiment using the same program and general materials (i.e., developed in PsychoPy; Peirce, [2007](https://www.tandfonline.com/doi/full/10.1080/02699930903485076)) which will generate identically formatted raw data files across all sites. We will then collate these data files from all sites and analyze them centrally using a single set of R code and scripts. All materials and analytic files will be pre-registered before data collection begins (see <https://osf.io/hs32y/)>.”

Second, to further test the fidelity of our experimental manipulation, we will conduct heterogeneity tests to examine if EC effects differ as a function of lab site (see p.14). Third, we adopt a random effects meta-analysis model (specifically, using the Restricted Maximum Likelihood method). We opted for this type of analysis as it account for possible differences between labs (see changes on p.14):

“Although all participating labs will use similar materials, differences may be introduced by the translation of materials, selection of stimuli, or characteristics of the samples. In order to account for this within the analyses, we will employ random effects meta-analysis models (specifically, using the Restricted Maximum Likelihood method)”

“The EC effect size in this group ranged from X.XX to X.XX across labs (see Figure X). The differences in EC effect sizes across labs were [consistent/inconsistent] with what one would expect by chance, τ = X.XX, *I*2 = X.XX%, *H*2 = X.XX, Q(X) = X.XX, *p* = .XXX”

**Reviewer 1**: If the proposed study is not the best way to conduct and/or analyze such a replication in what ways should it be improved?

N/A

**Reviewer 1**: **[R1.3]** What if any criteria should be established a priori on the publishability of the study (e.g., manipulation checks, off floor/ceiling, etc., measurement reliability)?

A protocol is needed to handle language and translation (presuming materials will be translated into locally dominant languages), and if necessary, subject recruitment (if all materials will be presented in English). Other known issues that can crop up in multi-national research (e.g., measurement equivalence) should be addressed as well.

**Authors**: Please see our response to R1.2.

**REVIEWER 2’S COMMENTS**

**Reviewer 2**: This registered replication report aims to establish how reproducible the evaluative conditioning effect is using Olson and Fazio's surveillance procedure. It is a collaborative effort across 12 labs at 10 universities in the US and Europe. The original authors provided input into the study design and one of the original authors is a co-author on this report.

In general, I find the study design to be a fair recreation of the original procedure. The sample size will provide high power to not only test significance but also to examine effect sizes and confidence intervals, and to perform a meaningful analysis examining variation in effect sizes across labs. The pre-registered analytic plan is good. I had only a few relatively minor questions/comments.

**Reviewer 2**: **[R2.1]** The target items (Pokemon characters) have been pre-tested to create a set of 20 affectively neutral and unfamiliar targets. Each lab will then select nine that are the most neutral and least familiar locally. This introduces some variability in the stimuli used across labs, although the items will still be sampled from the same set. Will the analyses treat target items as random effects (thereby treating them as samples from a larger population)?

**Authors**: In order to replicate the analytic strategy of the original publication, our unit of analysis is the standardized effect size at the laboratory level, which is then meta analyzed. As such, we cannot include a random intercept/slope for each specific stimulus in our confirmatory analyses, as this would require us to perform an Individual Participant Data Meta-Analysis instead. This change in our confirmatory analytic strategy and type of effect size would greatly reduce the reader’s ability to make direct comparisons between the results of the original study and our own replication.

Even if we were to do as Reviewer 2 suggests, and include an exploratory analysis of our findings that also “treats target items as random effects”, this would likely be unpromising. At best each lab will sample from 20 Pokémon characters that are severely restricted in their variance (i.e., low in familiarity and neutral in their valence). Conceptually, it is questionable whether aiming to generalize to a larger population of stimuli (by including a random effect) is a meaningful goal in this context. Statistically, the low number of CSs limits the power of the crossed random-effects model (see Judd, Westfall, & Kenny, 2017; p.17-19). We therefore believe that a cross-random effects analysis would only make sense if we increase the overall number of CSs. Yet, as we previously mentioned, this is not possible given that many labs will not pre-test stimuli for themselves (due to time and resource limitations) and will therefore rely on the two CSs selected from our (overall) pre-test.

Nevertheless, and as noted in response R1.2, we do use a random-effects meta-analytic model, which assumes that not all laboratories are necessarily studying the same effect. This is now clarified in the revised manuscript on p.14 with the following text:

“Although all participating labs will use similar materials, differences may be introduced by the translation of materials, selection of stimuli, or characteristics of the samples. In order to account for this within the analyses, we will employ random effects meta-analysis models (specifically, using the Restricted Maximum Likelihood method)”

**Reviewer 2**: **[R2.2]** In the surveillance task 9 of the ten affective images in each set are said to be the same as the original, but no explanation is given for why one image in each set is different. Please explain.

**Authors**: please see our response to Reviewer 1 (see R1.2) and changes in the revised manuscript on pp. 5-7.

**Reviewer 2**: **[R2.3]** In the exclusion criteria it is said that subjects will be removed if their error rate on the task is 3 SD higher or lower than the sample mean. It seems that only high error rates are a reason to drop subjects, not extremely low error rates (perfect accuracy).

**Authors:** We thank Reviewer 2 for this suggestion. However, given that this is a direct-replication report, we need to stick to the same criteria as utilized by the original authors. With regards to exclusions, this requires us to do as the original authors did, namely, remove subjects if their error rate on the surveillance task is 3 SD higher or lower than the sample mean.

**Reviewer 2**: **[R2.4]** Most importantly, I think the authors should test an additional hypothesis about the importance of contingency awareness to the evaluative conditioning effect. The proposal outlines several tests examining three different criteria for establishing (un)awareness. But in all cases, contingency-aware subjects are excluded. A plausible outcome of the replication is that the conditioning effect is reliable, but it is driven (mostly or entirely) by aware subjects. For theoretical and practical uses, it would be good to know if the surveillance task was an effective conditioning procedure when subjects are aware of pairings. So I suggest testing whether awareness (measured by each criterion) moderates the conditioning effect.

**Authors**: We thank Reviewer 2 for this suggestion. The revised manuscript now includes a

new set of exploratory analyses along these lines (see changes on pp.13,16-17). Specifically, for each laboratory, we will compute two EC effect sizes: one for participants coded as ‘aware’ and another for participants coded as ‘unaware’. We will then conduct moderator meta-analyses on these two scores to determine if they differ from one another. To account for the dependencies between effect sizes coming from the same experimental setting, we will add a random intercept at the laboratory level. Note that, in our exploratory analyses, we assess contingency awareness using four different criteria. We will therefore report the results of four moderator meta-analyses, each of them based on a different awareness criterion.

We would also like to acknowledge that, although these additional exploratory analyses are being included in our revised manuscript, several co-authors argue that such an analytic approach is not without its limitations. First, any attempt to detect differences in EC effects between putatively ‘aware’ and ‘unaware’ participants will ultimately depend on the reliability of the awareness measure and of the EC procedure itself. Yet, based on previous studies on similar unconscious learning effects (e.g., Kaufman, Deyoung, Gray, Jiménez, Brown, & Mackintosh, 2010; Vadillo, Linssen, Orgaz, Parsons, & Shanks, in press), we can’t be sure that the reliability of the awareness measures will be sufficient to detect a true difference between aware and unaware participants, even if such a difference exists. Second, it is conceptually and statistically problematic to use one outcome measure ("awareness") as a moderator of another outcome measure (EC). Statistically, both evaluation (i.e., EC effect) and memory performance (i.e., “awareness”) are measured, which implies that any relation between them is merely correlational and not causal (e.g., Gawronski &  Walther, 2012). Conceptually, the correlational nature of the relation between the two outcome measures make the particular relation between them ambiguous. For example, if we will find a difference between ‘aware’ and ‘unaware’ participants, that difference can reflect a moderating role of contingency awareness during encoding of the evaluative information. Alternatively, the same findings can be interpreted as reflecting a mediating role of the mental representation of the evaluative information (e.g., Gawronski &  Walther, 2012). The same logic holds in cases where we do not find a difference between ‘aware’ and ‘unaware’ participants. Because any possible outcome can be explained in at least two different ways, it is impossible to accurately interpret the results of the suggested analysis.

We now include a footnote highlighting these issues in the revised manuscript (see pp.13-14) and plan to address it in greater detail in the General Discussion.

“Note that the results obtained from such a comparison should be interpreted with extreme caution. First, any attempt to detect differences in EC effects between putatively ‘aware’ and ‘unaware’ participants will ultimately depend on the reliability of the awareness measure used, and of the EC procedure itself. Previous evidence suggests that unconscious learning paradigms and awareness tests tend to yield unreliable measures (e.g., Vadillo, Linssen, Orgaz, Parsons, & Shanks, in press). Second, it is conceptually and statistically problematic to use one outcome measure as a moderator of another outcome measure, due to the correlational nature of their relation (e.g., Gawronski & Walther, 2012). We will unpack both issues in greater detail in the General Discussion.”

**References**

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.‏

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.‏

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601-625.‏

Kaufman, S. B., Deyoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition, 116,* 321–340.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.‏

Vadillo, M. A., Linssen, D., Orgaz, C., Parsons, S., & Shanks, D. R. (in press). Unconscious or underpowered? Probabilistic cuing of visual attention. *Journal of Experimental Psychology: General.*