Dear Prof. Lindsay,

Please find attached the revised version of MS number PSCI-19-0402.R2 titled “Incidental Attitude Formation via the Surveillance Task: A Registered Replication Report 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)

**Response to the Review**

**Editor:**

Dear Ms. Moran:  
  
I hope that all is well with you and yours despite the challenges of these strange and stressful times.  I thank you and your team for submitting the Stage 2 version of "Incidental Attitude Formation via the Surveillance Task: A Registered Replication of Olson and Fazio (2001)" [PSCI-19-0402.R2] to Psychological Science. Thanks also to Reviewers 1 and 2 of the Stage 1 version, who gracious provided reviews of this one as well.   
  
This project is an impressive accomplishment and in my view warrants publication in Psychological Science. R1 recommended acceptance.  R2 recommended minor revisions that are quite manageable.  Bravo and congratulations. That said, in my view the paper needs quite a bit of revision to make it a good fit for this journal and to enhance the likelihood that it will be read and have the impact it deserves. Most of the hard work will be in making the paper shorter and more accessible to non-specialists.  It may not be easy but I am sure you can do it.  
  
**E.1.** The task description in the abstract and on page 5 is vague. In the latter, the first sentence turns out not to be strictly true and in any case it is a fine point.  Very early in the paper, please include a clear explanation of the gist of the procedure. Something along the lines of:  
  
“The CS stimuli were two Pokemon selected to be neutral and unfamiliar. The US stimuli were pictures and words, 4 of each selected to be positive and 4 negative.  In the training phase, on each of many trials, one or two stimuli were briefly presented, and subjects were told to monitor for a pre-designated target Pokemon (the “surveillance task,” a cover story that required subjects to engage with the stimuli). Most trials presented filler stimuli, but on interspersed training trials one of the two CS Pokemon was presented, always alongside a negative word/image whereas the other CS Pokeman was always presented with a positive word/image. Later, preference for the two CS Pokemon was assessed, followed by retrospective measures of awareness of the contingency during the training phase. The hypothesis was that subjects tend to prefer the CS Pokemon that had been paired with positive words/images, even if they report no awareness of the training contingency.”   
  
That is probably imperfect, but I hope you get the idea that it is crucial, early in the paper, to give readers who are unfamiliar with this literature a solid grasp of the procedure and the hypothesis. If they don’t have that within a page they will very likely stop reading.

**Response E.1.** In the revision, we included a clear explanation of the ‘surveillance’ task procedure (p. 2):

“In this task, commonly called the ‘surveillance task’, neutral and valenced stimuli are surreptitiously paired while the participants complete an unrelated task. Two neutral and unfamiliar Pokémon are selected to serve as conditioned stimuli. Valenced pictures and words serve as unconditioned stimuli. Participants are told that they will take part in a ‘surveillance task’ wherein they have to detect several target Pokémon that are different to the actual Pokémon of interest (i.e., the CSs) and press a key when they see them. During the task participants encounter many trials, some of which present a target Pokémon to which they have to respond, and others present (‘distractor’) stimuli to which they do not need to respond. Unbeknownst to them, several of the ‘distractor’ trials present CS-US pairs. Specifically, on some of the ‘distractor’ trials, one Pokémon (CS1) is always presented alongside a positive word or image (US positive) whereas on other ‘distractor’ trials a second Pokémon (CS2) is always presented with a negative word or image (US negative). In this way, the task requires people to process the CS-US pairs but directs their attention away from those pairings and towards the irrelevant target stimuli. Afterwards, relative preferences for CS1 and CS2 is assessed, followed by retrospective measures of awareness of the CS-US contingencies that were present during the surveillance task. Researchers who use this task assume that people will prefer CS1 (i.e., the Pokémon paired with positive stimuli) over CS2 (i.e., the Pokémon paired with negative stimuli), even if they later report no awareness of the CS-US contingencies (e.g., Jones et al., 2009, 2010; March et al., 2018).”

**E.2.** There are too many fine-grain details in the body of the text. I appreciate that you are striving to be transparent, but it is difficult to see the forest when so many trees are described in such detail.  The details are ready to hand for those who need them. For example, the info that two labs tested more than 150 and one tested less than 100 should be relegated to SOM.  The description of how Pokémon were selected could be made much more concise without loss.  There is no need to quote the verbatim instructions for the surveillance task in the body of the manuscript.  Steps taken to ensure fidelity and consistency across labs are admirable but need not be detailed in the main text – just point to them.

**Response E.2.** With this suggestion in mind, we removed as much fine-grain detail as possible from the body of the manuscript to the SOM. For instance:

1. We moved information about the labs who tested less than 100 participants or more than 150 participants to the SOM-R (p. 1, SOM-R).
2. We shorten the description of how the Pokémon were selected, and moved most of the details to the SOM-R (see p. 7 in the manuscript, and pp. 1-2 in the SOM-R).
3. We removed the quotation of the instructions for the surveillance task and the evaluation task from the body of the manuscript, and instead direct the readers to find these in the study protocol (see pp. 8, 9).
4. We shortened the section on experimental fidelity (see p. 10).
5. We removed a large amount of redundant material from the Results section, or moved it to the SOM-R or the discussion section, as appropriate (see pp. 11-17). The reporting of the preregistered analyses were not affected by this.

**E.3.** The exposition of the different exclusion criteria goes on for pages.  I recommend that in the text you explain the exclusion criteria used in the target article and then say “As detailed in SOM, we also analyzed our data with three alternative exclusion rules that categorized subjects as unaware of the contingency if they  (a) did not mention a systematic pairing between CSs and USs; (b) selected “No, I did not notice if that happened in my task;” and (c) did not confidently identify the correct CS that was paired with the US.” Compared to Olson and Fazio's criteria, these rules identified a larger percentage of subjects as "aware" of the continency."

**Response E.3.** In the revision, we moved the detailed description of the secondary exclusion criteria to the SOM-R (see pp. 2-4 in the SOM-R). In line with the editor suggestion, the main text now explains the secondary exclusion criteria in brief (see pp. 12-13):

“The original authors’ criterion may have led individuals who were aware to be scored as if they were ‘unaware’. We therefore preregistered three additional exclusion criteria to examine if evidence for EC effects in this task were robust to, or depended on, the specific way in which contingency awareness/recollective memory was measured. As detailed in SOM-R, the three alternative exclusion rules categorized participants as ‘aware’ if they: (a) referred to any form of systematic pairing between the CS and US stimuli (Olson & Fazio 2001 modified criterion); (b) indicated that one CS was systematically paired with positive USs and a second CS was paired with negative USs (Bar-Anan et al. 2010 criterion); or (c) in addition to (b) also correctly identified the valence of the USs with which each of the two CSs appeared (Bar-Anan et al. 2010 modified criterion). Compared to Olson and Fazio's original criteria, these awareness criteria categorized a larger percentage of participants as ‘aware’ of the CS-US contingency.”

**E.4.** You have a lot of discussion in your Results section. As explained in the submission guidelines, that is a no-no. Psych Science eliminated word counts from Results so that authors can report all of the analysis that are needed for an honest report of the findings, not so that they could include more discussion.

**Response E.4.** In line with the Editor’s request, we have removed unnecessary discussion from the Results section and focused solely on the analyses themselves (see revised Results section on pp.11-17). Some points were reduced in length and merged into in the discussion section (see p. 18, second paragraph of discussion).

**E.5.** I know you already have a detailed document address deviations from the preregistered research plan, but please also complete and submit the report from the Transparency Checklist <https://www.nature.com/articles/s41562-019-0772-6R2>

**Response E.5.** We have completed the Transparency Checklist and included it along with our resubmission and on our OSF project page ([osf.io/8tmfv/](https://osf.io/8tmfv/)).

**E.6.** R2 makes the astute point that “retrospective reports of awareness create uncertainty in both directions – they may classify unaware subjects as aware, but they may also classify aware subjects as unaware.”  I think your revision should mention that point.  I also agree with Rs that “the new data raise doubts, but do not settle the question of whether evaluative condition can happen without awareness.”

**Response E.6.** In line with Reviewer 2’s comments we now acknowledge that retrospective reports of awareness create uncertainty in both directions and that the data generated in this study raise doubts but do not settle the question of whether EC can happen in the absence of awareness (see changes in the Discussion on pp. 19-20):

“Although our primary analysis demonstrated that Olson and Fazio’s (2001) surveillance task effect was replicated, these conceptual concerns raise questions as to whether this procedure represents a useful test of the ‘unaware’ EC hypothesis. Retrospective reports of awareness are imperfect in that they may misclassify participants as unaware or vice-versa (but see Hussey & Hughes, 2020). Nonetheless, data based on retrospective measures, such as those used here, likely cannot settle the question of whether EC effects can emerge in the absence of awareness by themselves.”

**E.7.** I think it would be helpful to add to your OSF project page an overall guide to your OSF materials -- I found it a difficult to navigate. So a TOC might be helpful.

**Response E.7.** We have added a Table of Contents for our OSF project (see [osf.io/nf7vh/](https://osf.io/nf7vh/)).

**E.8.** I am sorry but I am not open to publishing Commentaries from various subgroups of authors.  As an alternative, I propose that we explore the possibility of placing registered copies of brief statements from those who wish to tender them on the OSF page for the project, and including in the article the url for the collection of statements.  Would that be satisfactory?

**Response E.8.** We thank the Editor for considering our proposal. In light of his decision we decided to create an OSF page for the various commentaries (see [osf.io/qtcsw](https://osf.io/qtcsw/)) and to cite these commentaries in the manuscript (see pp. 18-19). Should additional commentaries emerge in future we will also collate them there.

Very minor matters  
  
**E.9.** How about changing the title so that it says “…a Registered Replication Report…”?  I am not wild on these terms, but that’s what this sort of project (multiple labs all following the same protocol) are currently being called.  A “Registered Report” is a slightly different beast.

**Response E.9.** We have now changed the title in line with the Editor’s suggestion (i.e., refer to it as a Registered Replication Report).

**E.10.** This is trivial, really just a pet peeve, but in my view most if not all occurrences of “actually” are actually not necessary.

**Response E.10.** In our revision, we have deleted the word “actually”.

In a separate communication, the editor had another request for revisions:

**E.11.** Greetings.  I heard recently from Michael Olson, copied here, asking if I would be open to considering a Commentary on the above-referenced manuscript.  As you know, in my recent action letter I indicated that I was not open to the idea of publishing Commentaries by coauthors of the piece, and instead floated the idea of including in the revision pointers to registered documents in which subsets of coauthors expressed their views. Those would, in content, be comprable to Commentary and Reply pieces, but would not be publishes as separate articles but instead be treated as part of the Supplemental Online Materials.

In considering Michael Olson's email, I have come to belive that there is a better approach. I am emailing you both to ask that you consider it.

As a preamble, I'll note that in my view all authors of a piece must be in accord with the core content of the piece. If some party to a piece of work is not willing to take on the message of that work then that party should decline to be listed as authors. That's what authorship means. In some such cases individuals might choose to be recognized in some other way (e.g., in the Author Note).

But in my view it is not necessary that an article in Psych Science give a single, unified message about the meaning of the results.  I therefore propose to you that you consider a Discussion section along the following lines:

"All coauthors of this work agree on the importance of the questions explored in this experiment and on the accuracy and completeness of the report of the methods and the results. But we were unable to come to consensus with regard to the interpretation of the findings.  Olson and Fazio, authors of the original article, believe that what is most important about the present findings is the evidence for an EC effect when 'aware' participants were excluded using the criteria used in the original study. That is a successful replication of the original effect.  [some other sentences articulating Olson and Fazio's views regarding the implications of the results, perhaps running to a paragraph or two]. The remaining authors, in contrast, believe that what is most striking about the results is the tenuousness of the support they offer for the core claim. [elaboration of that perspective]."

Obviously what I have offered here is merely a brief sketch, but I trust that you get the basic idea, which is that it is not necessary (or perhaps even desirable) for the article to take a unified stance on what the results mean. I believe you are all people of good will, and I am hoping that this approach will enable you to craft an approach that enables you to stand behind the report or the work. If you need a hundred or so extra words to do that, so be it.

**Response E.11.** In our revision, we solicited and included the original authors perspective on the results. We clearly label their response as being authored by them and that we include their text verbatim. Prof. Olson has also confirmed by email that he is satisfied with this section and its location in the manuscript. These can be found starting on page 20:

“A brief response was solicited from the original authors and we include it here verbatim. “We [Olson and Fazio] emphasize that the effect was in the predicted direction in 11 of the 12 samples using the original exclusion criteria. The secondary criteria revealed analogous patterns in 10, 9, and 11 of 12 samples, respectively. However, such criteria can also exclude unaware individuals if they use their recently formed attitudes to guess CS-US valence (see Gawronski & Walther, 2012). Ultimately, the lack of a moderating effect of exclusion criteria can be interpreted as an unqualified replication of Olson and Fazio (2001). In addition, the effect size produced by a single procedure is minimally relevant to broader theoretical questions about the multiple mechanisms that produce EC. Within our proposed implicit misattribution mechanism, the magnitude of EC is dependent upon source confusability (the extent to which the evaluation evoked by the US is likely to be misattributed to the CS; Jones et al., 2010). Hence, future work should focus on fostering source confusability beyond the procedural parameters employed here.”

**Reviewer: 2**

Comments to the Author  
This registered replication report tested the evaluative conditioning paradigm introduced by Olson & Fazio (2001), using four different criteria for measuring awareness of the pairings. Overall, the conditioning paradigm resulted in conditioned evaluations. However, conclusions of conditioning without contingency awareness depended on the criteria used for awareness. When the original authors’ criterion was used, a small but significant conditioning effect was observed. The other three criteria, which all excluded substantially more participants as “aware” led to no significant conditioning effect. The difference between criteria, however, was not significant. The authors conclude that although the conditioning effect was replicated, there is not strong evidence for the hypothesis that conditioning occurs without awareness of the pairings.  
  
This project appears well designed and executed, and the study is well-powered. The reported analyses are consistent with the pre-registered plans (and other analyses are labelled as such). In general, I believe the conclusions drawn are appropriate based on the findings. However, there are three points that I believe the authors should acknowledge in interpreting the results.  
  
**R2.1.** First, the authors acknowledge that retrospective reports of awareness can be problematic. They note that one problem is that participants may be aware of the pairing during the surveillance task but not remember accurately on the awareness check (thus underestimating awareness). But they should also note that participants may be unaware of the pairings during the surveillance task and then make an inference during the awareness check based on their evaluations. This would lead to an overestimate of awareness. This is especially likely using the Bar-Anan method, which prompts subjects to consider the pairings and then gives them a multiple choice test. Taking subjects at their word as to whether they are guessing in this context is not useful, as it assumes near perfect meta-cognitive knowledge. So, retrospective reports of awareness create uncertainty in both directions – they may classify unaware subjects as aware, but they may also classify aware subjects as unaware.

**Response R2.1.** In line with Reviewer 2’s comments we now acknowledge that retrospective reports of awareness create uncertainty in both directions (see changes to the Discussion on pp. 19-20):

“Further reasons for caution can be found in the ‘awareness’ concept itself. Debate continues to rage about what such exclusion criteria actually capture: some argue that it is ‘awareness’ (Jones et al., 2009) whereas others advocate for ‘recollective memory’ (Gawronski & Walther, 2012). For example, participants may be aware of pairings during the acquisition (EC) phase but fail to recall this information during the retrieval (evaluative) phase. Although our primary analysis demonstrated that Olson and Fazio’s (2001) surveillance task effect was replicated, these conceptual concerns raise questions as to whether this procedure represents a useful test of the ‘unaware’ EC hypothesis. Retrospective reports of awareness are imperfect in that they may misclassify participants as unaware or vice-versa (but see Hussey & Hughes, 2020). Nonetheless, data based on retrospective measures, such as those used here, likely cannot settle the question of whether EC effects can emerge in the absence of awareness by themselves. Alternative experimental manipulations of awareness are also possible, however results from such studies also fail to produce consistent evidence of ‘unaware’ EC (e.g., Corneille & Stahl, 2019).”

**R2.2.** Second, the moderation results are informative. They suggest that as larger and larger segments of the sample are classified as “aware,” the effect size in the segment classified as “unaware” is reduced. In other words, there is a correlation between self-reported awareness of the pairings and the size of the evaluative conditioning effect. This is the same as has been found in other evaluative conditioning paradigms. The authors urge “extreme caution” when interpreting the moderation effects because “it is conceptually and statistically problematic to use one outcome measure as a moderator of another outcome measure, due to the correlational nature of the finding.” This is true. But the problem is \*identical\* for using one outcome measure as an exclusion criterion when looking at another outcome measure. Excluding people based on the awareness check suffers from the same problem of dependency as the moderation analysis does (for the reasons described in point 1 above). Put differently, reporting and interpreting the results for only the “unaware” group amounts to reporting only one simple effect without acknowledging the interaction or the other simple effect (the “aware” group).

**Response R2.2.** We agree with the reviewer’s point here, however, this is a criticism of the original study’s methodology rather than our replication of it. Our replication of the methods used in the original study are not an uncritical endorsement of these methods. As we discuss in our introduction (p. 3), the original study is highly cited despite a number of factors that may suggest that its results may not fully address the underlying research question of ‘unaware’ EC. Our RRR examined the replicability of this effect, and our discussion section also critiques the methodology’s ability to address the original research question and points to alternative approaches such as the experimental manipulation of awareness (see p. 20: “Alternative experimental manipulations of awareness are also possible, however results from such studies also fail to produce consistent evidence of ‘unaware’ EC (e.g., Corneille & Stahl, 2019).”). The discussion section also references two commentaries papers by some of the RRR authors that address the reviewer’s points (see pp. 18-19). First, Stahl and Corneille (2020), account for the dependency between the ‘aware’ and ‘unaware’ groups by distinguishing between independent sets of fully ‘unaware’, partially ‘aware’, and fully ‘aware’ participants, and show that evaluative learning effects are moderated by participants’ set and are found in the latter two sets only. Second, Hussey and Hughes (2020), consider in detail the distinction between a replicable effect (as we found in our primary analyses) versus support for the underlying verbal hypothesis of interest (i.e., ‘unaware’ EC). Hussey & Hughes’s commentary supports and agrees with Reviewer 2’s point here that awareness rates are correlated with the EC effect in the supposedly ‘unaware’ sample. We have therefore caveated our results and conclusions to consider the issues that Reviewer 2 raises in the ways we have mentioned here, as well as by continuing this discussion at length in commentary pieces. Finally, we have removed the quote that Reviewer 2 referred to in order to remove any potential for confusion that we considered this issue to be specific to the moderation analysis. This section had already been moved to the SOM-R on the editor’s request (see above).

**R2.3.** Finally, the pattern of results suggests that there is a correlation between reported awareness and the size of the conditioning effect. The different criteria are essentially different ways of drawing the line between “zero awareness” and “some awareness.” In other words, there is a slope, but where is the intercept? So I think the paper is correct to conclude that the new data raise doubts, but do not settle the question of whether evaluative condition can happen without awareness. These problems all arise from using a retrospective report of awareness, which I believe is the biggest methodological lesson from the project.

**Response R2.3.** We agree,no single paper can or should conclusively decide a debate. Our conclusion section emphasizes that doubts have been raised and strong evidence for unaware EC is lacking, but does not conclude that it is a settled debate (p. 20):

“To conclude, although we replicated the surveillance task effect, we urge caution when using such an effect to make strong claims about ‘unaware’ EC, especially when those claims are being used to justify new theory or interventions. We also encourage more careful reflection on existing theory and interventions that have already been founded on this effect (e.g., March et al., 2018; Shaw et al., 2016). Strong claims necessitate strong evidence; evidence that is currently lacking.”

**Reviewer: 1**

Comments to the Author  
**R1.1.** The authors for the most part adhered closely to their preregistered plans. Where they did not, they faithfully documented their reasons for deviation in a supplemental online file.  
  
I carefully read their explanations for deviations and was persuaded they were reasonable. The choice to shift their language from confirmatory and exploratory to primary versus secondary made sense to me, given exploratory usually implies that the analyses are post hoc rather than planned a priori. Given the authors registered their hypotheses and planned analyses, this change in language helped to remove some potential ambiguity.  
  
The authors also preregistered that they would report an "overall EC effect," not realizing that a multilevel moderator meta-analysis model does not yield an overall effect (and in this case, it yielded four EC effects). I think their choice of how to handle this was appropriate, and as noted, it did not change their preregistered plan for deciding whether the effect under consideration replicated.  
  
The authors also explained a deviation in planned sample sizes and their data collection stopping rule, which had implications for analysis. The reason for the deviation is understandable and was dealt with in a fair and reasonable way.  
  
Taken together, I am satisfied that the replication held sufficiently true to the registered plan

**Response R1.1.** We thank the reviewer for their positive feedback.