**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 tht 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 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 (pp. 5-6):

“In this task, commonly called the ‘surveillance procedure’, the CS stimuli are two Pokémon, selected to be neutral and unfamiliar. The US stimuli are valenced pictures and words. Before the conditioning phase, participants are given a “surveillance task,” cover story that asks them to detect and respond to target Pokémon (which are different from the CS stimuli). The conditioning phase includes many trials, some of them present the target Pokémon to which participants need to respond, but most of them present other (“distractor”) stimuli. Unbeknownst to the participants, several of the “distractor” trials present CS-US pairs. In other words, in some of the “distractor” trials, one Pokémon is always presented alongside a negative word/image whereas in other “distractor” trials the other CS Pokémon is always presented with a positive word/image. In this way, the task requires people to process the CS-US pairs but directs their attention away from those pairings and towards irrelevant target items (Jones et al., 2010). After the conditioning phase, preference for the two CS Pokémon is assessed, followed by retrospective measures of awareness of the contingency during the conditioning phase. The main hypothesis underline this task that participants tend to prefer the CS Pokémon that had been paired with positive words/images, even if they report no awareness of the CS-US contingency (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 Pokemon 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.** In the revision, we removed some of the fine-grain details from the body of the manuscript.

Specifically:

We moved the information about the labs who tested less than 100 participants or more than 150 participants to the SOM-R (p. 9 SOM-R).

We shorten the description of how the Pokémon were selected, and moved most of the details to the SOM-R (see p. 11 in the manuscript, and p. 9 in the SOM-R).

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. 12, 14)

We shortened the experimental fidelity section (see p. 16).

**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 p. 10 in the SOM-R). In line with the editor suggestion, the main text now explain the secondary exclusion criteria shortly (p. 19)

“We considered that the original authors’ criterion may have scored individuals who were actually aware of/remembered the contingencies as ‘unaware’. Therefore we preregistered three additional secondary exclusion criteria that allowed us 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, the three alternative exclusion rules categorized subjects as aware of the contingency if they (a) referred to any form of systematic pairing between the CS and US stimuli (Olson & Fazio 2001 modified criterion); (b) selected “Yes, that happened in my task” on question 1 from the Bar-Anan et al. protocol (Bar-Anan et al. 2010 criterion); (c) in addition to (b) also correctly identify (on questions 2 and 3 from the Bar-Anan et al. protocol) 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 criteria, these rules identified a larger percentage of subjects as "aware" of the continency.”

**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.**

**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.**

**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.**

**E.7.** I think it would be hlepul 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 added to our OSF project an overall guide to the OSF materials (see LINK).

**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.**

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.**

**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 the revision, we deleted most of the occurrences of “actually”.

**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.**

**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.**

**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.**

**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.