Dear Dr. Hussey,

Thank you for submitting your manuscript for consideration by Collabra: Psychology. I have now received two reviews of your manuscript from researchers with considerable expertise in your field. These reviewers provided expert feedback and I thank them for their service to this journal. I also independently read your manuscript before consulting these reviews. As you will see, the reviewers differed quite dramatically in their assessment of the manuscript. One reviewer was quite favorable to the work and another reviewer was quite unfavorable. I find my views to be very similar to the views of the reviewer who was unfavorable in their assessment of the manuscript. Thus, I am rejecting this manuscript from Collabra: Psychology. I am sorry to have to bear this disappointing news. Below I note the issue that factored most heavily in my decision.

To me, the fundamental issue present in this paper is the interpretation of the root data. Based on the extant literature (much of which one of the reviewers points to) the concern would be that the participants who failed the awareness check as representing participants that should be discarded from the data analysis entirely. Certainly, much (if not most) of the research looking at online data collection suggests that the participants should be dropped from analyses entirely as they haven’t likely been paying attention. This has been noted in mTurk samples (Aruguete et al 2019) as well in more lab traditional sources (Meade & Craig, 2012). Importantly, by most metrics, the participants you are focusing on should simply be dropped as careless participants (e.g., Huang, et al. 2012, and 2015, Meyer et al 2013, Weathers & Bardakci, 2015; you could also look at Nichols & Edlund, in press).

In closing, I hope that these reviews are helpful as you pursue your work. Both reviewers provided excellent reviews that I believe demonstrate the best of peer review and how it is supposed to work. I also hope that this negative editorial decision does not deter you from considering Collabra: Psychology as a future outlet for your research.

Sincerely,

John E. Edlund

**Reviewer 1**

This article presents a reanalysis of data reported in Moran et al. (2020), which is a multi-site registered replication report of data collected using the Olsen and Fazio “classical conditioning” paradigm along with alternative data treatments. Moran et al. replicated the original Olsen and Fazio effect, and did so when using the same exclusion criterion employed in the original paper. The Olsen and Fazio effect was not replicated when each of three other data exclusion methods were employed. The current manuscript adopted a conditional Boolean approach to data exclusion, retaining data only if examination simultaneously passed all 4 exclusion criteria. The authors of the manuscript under consideration see their new analysis as important enough to regret not including it in the original analysis plan. They suggest that this new analysis would “have increased the evidential weight of the current results.” Even if I agreed that the manuscript authors’ additional analysis was correct, I’m not sure that I agree with the authors’ conclusion that their new analysis results justify publication. Here’s why.  
The original Moran et al. manuscript results show that the Olsen and Fazio effect replicates when their original exclusion criteria were used [participants were scored as ‘aware’ and were discarded if they wrote that CSpos (either its name or a description of its appearance) appeared during the task together with positively valenced words/images and that CSneg (its name or a description of its appearance) appeared during the task together with negative words/images]. The effect did NOT replicate with a more liberalized criterion [participants were coded as “aware” if they (a) identified only one of the two CS-US pairings, (b) identified that the two CS were paired with US stimuli but not specifying the specific way in which the CSs and USs were paired].  
I suspect that these two criteria are not independent – if I am thinking about this correctly, then logically if a subject was classified as “aware” via criterion 2 then they MUST also be classified as aware via criterion 1. Thus, when using a Boolean “or” decision rule, classifications produced by criterion 1 are subsumed by the classifications produced by criterion 2. In other words, in any Boolean ‘or’ decision rule, criterion 1 rejections are redundant with, and add no new information to, the rejections made by criterion 2 (criterion 2 can be seen as all the criterion 1 rejections, plus some more). So why include criterion 1 in the joint decision rule? I don’t get it.

Moreover, I’m not smart enough to definitively see whether it is appropriate to include such items in an assessment of the extent to which such items conform to a Guttman structure. My intuition says “no.” My initial intuition is also that Mokken modeling for these data might be preferable to the basic Guttman approach. The deterministic Guttman scaling model unrealistically assumes that the data are error-free. Mokken models bring the Guttman idea within a probabilistic framework and therefore allow researchers to model data allowing for measurement error.

However, I admit my intuition may be wrong. Important to this approach is that there is an assumption of unidimensionality, which means that all items from the same instrument, test or subscale i.e. the “item set” share (measure) the same latent trait (θ) apart from any unique characteristics of (any) item(s) and the presence of some ubiquitous measurement error. I’m not at all sure that the data justify this assumption (and neither the original paper not this manuscript show me correlations that might allow an independent assessment of construct unidimensionality – but the manuscript authors’ own statement that there is no “structural validity” evidence says that there may not be unidimensionality).

In addition, these kinds of IRT-style analyses routinely assume that there is independence among items. This assumption is that an individual’s response to an item is not influenced by his or her responses to the other items in the same test/scale. This might be violated by the Moran et al. procedure. Might the fact that the constant presentation order used in the procedure [“we used the original Olson and Fazio (2001) post-experiment questionnaire followed by the questionnaire used in the studies of Bar-Anan et al. (2010)”] have lead to order effects in responding to the Bar-Anan items? If true, that might mean that ONLY the Olsen and Fazio items were “pure” in that they were potentially uninfluenced by responses to other probe items.

Thus, one might be tempted to offer the argument was that it was because they came first in the procedure that the Olsen and Fazio items “worked,” and that responses to the Bar-Anan et al. items failed because people may have been “sensitized” to the procedure by the implications of the Olsen and Fazio items. This line of thought suggests that the original analysis results that came from the Bar-Anan items are to be looked on with some doubt. Given this level of doubt, the analysis produced by the authors of this manuscript, which uses the same potentially-dubious items, is similarly in doubt. It is true that the manuscript authors might think that this line of thought is highly fanciful because I have no evidence for it. But by the same token, I would similarly suggest that, given the data that they have in hand, the authors can produce no evidence that my argument is not valid.

Another reason to be skeptical of the authors’ analysis is that their use of the Boolean ‘or’ decision rule seems to imply that the 4 exclusion methods are all valid indicators of “unawareness”. One needs to consider this carefully. What the measures are after in these probes is evidence that at least some subjects attended to and processed the stimuli, but were influenced by the stimuli without realizing it. To my knowledge, no one in this area, including the manuscript authors, has taken the time to generally establish the psychometric validity of any of these post-procedure probes with respect to this “attended to but influence-unaware” idea.

Indeed, I can point to recent psychometric practices that suggest that deleting those who can accurately report critical information from the paradigm is exactly the WRONG thing to do. Many studies now include procedure checks in which subjects are asked to recognize or recollect critical information from the paradigm. If they cannot, it is THESE subjects who are discarded under the assumption that they not complying with instructions and/or were inattentive to the procedure. In such studies it is those subjects who recognized or recalled the critical information who are retained for the analysis, with the “noncompliant dregs” discarded. This IS EXACTLY THE OPPOSITE OF WHAT IS BEING DONE IN THESE Classical Conditioning studies. From the perspective of these “discard the dregs” advocates, reducing or losing the effect of interest is EXACTLY what should happen when one analyses data from those who did not provide sufficient evidence of being attentive to, and being engaged in, the procedure.

The key conceptual issue, then, is what does failure to recognize or recall critical stimuli mean? Does it mean that subjects were engaged in the procedure but were just so overwhelmed or overloaded that they “forgot”? Or does it mean that they just blew the procedure off (a common problem in research - see Meade & Bartholomew, 2012; McKibben & Silvia, 2015; Huang et al., 2014; Huang et al., 2015; Berry et al., 2019)? The authors don’t know (nor do those who did the original research, or the replications and extensions). I don’t know either, but at least I can point to a corpus of data that has emerged among those concerned with data quality suggesting validity in the discard of “noncompliant dregs”: When one deletes the “noncompliant dregs” because of lack of recognition, the data tends to improve.

My position becomes even more viable when one recognizes that the technique that was “gentlest” in discarding subjects in Moran et al. was the one that continues to produce effects. The original Olsen and Fazio probes are pretty vague at detecting attentiveness to the procedure and the standard for discarding people is pretty high. Thus, one might argue that the O&F procedure was least aggressive in discarding attentive subjects. One might argue that the other procedures were more sensitive in detecting and deleting attentive subjects, retaining only the “dregs”, which is why the effect dissipated using those methods.

In this regard, I wondered whether the magnitude of the effect observed was related to how many of the measures a discarded subjects “failed”. What was the magnitude of the effect in those discarded because they failed to pass one method? Two methods? Three methods? All 4 methods? I might wonder whether the classical conditioning effect was largest among those subjects who were most engaged in the procedure.

I was also puzzled by the fact that the manuscript authors argue that their analysis indicated a lack of both reliability and structural validity in the 4 methods used to discard subjects, yet they went ahead and used the 4 indicators as if they were both reliable and valid when applied in their joint-decision Boolean ‘or’ decision rule. By what logic is it correct to use psychometric items that don’t reliably and/or validly coalesce in this way? Any scaling procedure would argue that one should do additional things to “improve the measure”. This would include deleting bad items, and/or generating new items that provide a better fit with the items retained. I definitely require enlightenment as to why the authors thought that their use of the 4 variables in their Boolean decision rule was a psychometrically reasonable thing to do, because right now I don’t see that as a justifiable procedure.

Another gripe was sparked by the statement in the Moran et al. (2020) piece noted, one the manuscript authors repeated, the argument that “claims for the replicability of support for the verbal hypothesis of ‘unaware EC’ have far reaching implications, and such claims require strong evidence.” Given this statement, I was baffled by the fact that the manuscript authors wrote their manuscript in seeming ignorance of the wider contest in which classical conditioning studies have been used. It seems to me that evaluating the strength of the “unaware conditioning” idea requires one to extend beyond the Olson and Fazio paradigm. If one is really interested in the strength of the evidence related to the awareness” issue in evaluative learning, then one needs to look outside the restricted confines of the Olsen and Fazio paradigm. Indeed, this issue of awareness/unawareness shows up all over the place. The issue appears in the Hoffmann et al. (2010) meta-analysis, in the Mertens et al. (2019) meta-analysis of classically conditioned fear, and in work that supposedly explores the use of subliminally presented stimuli in classical conditioning (e.g., Raes & Rudi, 2011; Veltkamp et al., 2011; Milner et al., 2017; Amd and Baillet, 2019; Wagner et al., 2020). The issue also appears elsewhere in the attitude literature (e.g., Betsch et al., 2001). To view this manuscript appropriately, in my opinion readers need to be made aware of (or reminded of) the broad empirical and theoretical context in which the authors’ work resides.  
One important reason why this needs to be done lies in the fact that the order and timing of stimulus presentation differs across classical condition procedures. At least 4 methods are described: trace, delay, backward, and simultaneous. These procedures all don’t work in the same way (e.g., see Cautela, 1965; Mackintosh, 1974; Oristaglio et al., 2013; Heth & Rescorla, 1973; Dostalek, 1976; Spetch et al., 1981). Indeed, the simultaneous stimulus presentation method used in Olsen and Fazio is thought be many to produce relatively weak effects. Moreover, some have claimed that the different procedures produce conditioning effects for different theoretical reasons (see Durkovic & Damianopoulos, 1986; Wagner et al., 2020). Thus, while the authors write as if their conclusions apply to all of classical conditioning, they really apply only to the exact simultaneous conditioning procedure that is used in the Olsen and Fazio paradigm. Again, to make this point clear, the authors need to place their research squarely in the context of all the other classical conditioning stuff that is out there.

**Reviewer 2**

To the authors, I had the pleasure of reviewing your manuscript “Evaluative Conditioning without awareness: Replicable effects do not equate replicable inferences.” In it, you discuss the issues of mapping statistical effects onto verbal hypotheses in the case of evaluative conditioning. Overall, I really like this paper. I think it’s a great demonstration of the nuances involved in replication. Hopefully, it will serve as a useful example for researchers when trying to evaluate replications. I’ve got a couple of suggestions, mainly places to add more detail (if possible), but nothing that calls the core of the paper into question. The section on Guttman errors seems really important for demonstrating the issues with the awareness checks. If you’ve got more space to work with (this looks like a commentary, so maybe you’re up against length requirements), I’d expand your explanation a bit more. As long as I’m understanding correctly, the idea here is that if they’re perfectly reliable, the people who failed various checks should be nested within each bucket as the tests get stricter. So basically, anyone who fails a more lenient test should also fail the ones that are stricter. If it’s not too difficult, it might be really useful to create a figure here to demonstrate this principle. Again, if I’m understanding correctly, if you made a Venn diagram of the 4 criteria, a Guttman error of 0 would include four concentric circles (moving from most lenient to strictest), a score of 1 would be 4 non-overlapping circles. What you observe seems to be about in the middle of that spectrum. Visualizing that might be really useful for readers. Another place I’d expand a bit on your thinking is in describing the rationale for combining the exclusion criteria. I think one response to your strategy might be, well you just made the case that these exclusion criteria aren’t great for assessing awareness, why are you using them now? Again, my understanding here is that while each measure is flawed, each measure is capturing some bit of awareness. Although we’re probably making some errors along the way in using them all, to the extent that some people are incorrectly failing the awareness checks, the sample is big enough that we can take that conservative approach and exclude them. If I’m understanding your reasoning here, I’d try to make a bit more explicit in the paper.

Warm Regards, Charlie Ebersole University of Virginia [cebersole@virginia.edu](mailto:cebersole@virginia.edu)