Re: PSP-A-2019-0728

The AMPeror’s New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes

Journal of Personality and Social Psychology: Attitudes and Social Cognition

Dear Mr. Cummins,

I have received two very thoughtful reviews of the manuscript that you and your co-authors recently submitted to JPSP-ASC, titled "The AMPeror’s New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes" (PSP-A-2019-0728). I am deeply grateful to the reviewers for the time and effort they put into their reviews, which were very helpful in reaching this decision. Furthermore, I read your paper carefully and independently, before looking at the reviews.

As you can see when you have had a chance to see the reviewer comments, we all find this line of inquiry to be promising and a valuable contribution to the field. The AMP is widely used, so we all appreciate the goal of taking a careful look at how it operates and the degree to which awareness of the primes influences the focal effect. At the same time, we all note substantial weaknesses with the study, particularly with regard to alternative explanations of your results and how well your paper is positioned to contribute meaningfully to previous conversations about the validity of the AMP. I am sorry to report that I cannot accept the current version of the paper for publication in JPSP-ASC. However, in recognition of the potential value of this line of work to researchers in the field, I would be willing to entertain a substantial revision (submitted as a new manuscript) if you believe it is possible that a revision would be able to address the reviewers' critiques.

**Authors**: We thank the Editor and Reviewers 1-2 for their constructive and helpful feedback. We took that feedback seriously when revising our manuscript. We also conducted two new studies that directly speak to several of the issues that were raised. We believe that our re-submission is considerably stronger as a result.

Briefly, there were three core arguments in our original paper. Reviewers 1 and 2 focused on two of these points and highlighted concerns that they had (i.e., issues to do with the proposed mechanism that mediates AMP effects and XXX). We offer a detailed reply to the Reviewers in our response letter (*see below*) and have revised the manuscript accordingly. We also draw attention to a third point that we feel was overlooked during the review process but which was perhaps the most important finding of our paper (i.e., whether the AMP effect is a valid measure of evaluation).

We briefly summarize these three points and our response here for your convenience.

**Issue 1: What is the operating process behind AMP effects**?

Very short intro to issue.

To address this concern (i.e., that the IA-AMP effect [and by implication standard AMP effects] were mediated by posthoc confabulation) we conducted two new studies. Here we further modified our IA-AMP so that participants had to indicate if they were aware of the prime’s influence on their target evaluation either (a) before the target evaluation was emitted (Experiment 6) or (b) before the target stimulus was even presented onscreen (Experiment 7). Under these conditions it is highly unlikely that the AMP effect is driven by the fact that an individual has generated a post hoc confabulation to explain their target evaluation, given that the target evaluation has yet to be overtly emitted (Experiment 6) or covertly generated (Experiment 7). Yet even under these conditions we still obtain similar effects as reported in Experiments 1-5.

In short, we now provide evidence from seven studies that IA-AMP effects are *not* driven by post-hoc confabulation, and instead driven by *a priori* awareness of the prime and its influence on one’s target evaluation. This *a priori* awareness of prime influence at Time 2 also predicts the magnitude of standard AMP effects at Time 1. Although we recognize that one could of course modify a post-hoc confabulation account to accommodate our findings, any such argument would itself be post-hoc. We are aware of no empirical arguments in the AMP literature or theorizing that would a priori predict the findings reported here. In contrast, our findings and predictions were consistently pre-registered. Thus, while recognizing the value of post-hoc theorizing we also recognize that it should not be afforded equal weight to *a priori* pre-registered hypotheses.

For more on this see Experiments 6-7 as well as the revised General Discussion

**Issue 2**. **What is the operating conditions under which the AMP effect is emitted?**

In short, XXX

For more on this see XXX

**Issue 3. Have others not already addressed this issue and have to mischaracterised their work?**

In short, XXX

For more on this see XXX

**Issue 4. Does the AMP effect represent a valid measure of evaluation**. Whereas Reviewers 1-2 focused on issues 1-2, we believe that a third critical issue was overlooked in our previous submission.

In short, XXX

For more on this see XXX

**Editor**: The reviewers clearly expressed their concerns and thus I will not reiterate them. However, let me highlight a few points that are most important. First, both reviewers pointed out specific ways that previous work (e.g., Payne et al., 2015 and Gawronski & Ye, 2015) was mischaracterized. Usually, miscommunicating details of the method of previous work would not be considered a major flaw of a paper, but it is quite troubling to see in a paper whose raison d'etre is to identify flaws in an experimental procedure used by other researchers. In this case, a fair scientific debate demands that the critic be accurate and specific about exactly what the flaws of the prior work are.

**Authors**: We thank the Editor and Reviewers 1-2 for highlighting this issue to us. We apologize for any mischaracterization – it was not our intent. In our original submission we focused on the issues that we perceived with the literature and failed to acknowledge additional details which were also relevant to the papers being discussed.

In-light of this feedback we have revised our manuscript to better represent the methods and findings of past work. Nevertheless, we still stand over our original core point that the literature contains a number of statistical, methodological, and conceptual issues, which we feel is only fair to acknowledge (see revisions on pp.XX). While we accept the points that Reviewers 1 and 2 made and have adjusted the manuscript accordingly, we also ask that these overarching points and their implications be engaged with fully.

**Editor**: Second, both reviewers also point out that mere awareness of the primes does not imply that the awareness "drives" the focal AMP effect. (As a side note, we all found the use of the term "drive" throughout to inappropriately imply a causal effect.) The reviewers are quite articulate about the issues here, with the overarching conclusion being that a causal effect of awareness is only one among many possible explanations for the AMP effect. The alternatives are not adequately ruled out and, as Reviewer 2 notes, all the observed effects in this paper are consistent with implicit misattribution plus some post-hoc justification. The evidence presented here simply does not support the conclusion that the AMP is invalid in the ways you claim (or even really in any ways), so even as the general undertaking of interrogating the AMP might be valuable, the specific way it was implemented here does not contribute much to the literature.

**Authors**: In line with the Editor and Reviewers 1-2’s suggestion we have amended any reference to the idea that prime influence awareness “drives” AMP effects in the revised manuscript. Our original intention in using the term drive was to convey that the majority of variance in AMP effects is made up by the minority of highly aware participants and the trails on which they were aware. However, in order to avoid the issue of the term being interpreted as implying causality we have substituted it for “is associated with”. Hopefully this is less contentious.

We think it is highly important to note, however, that our argument – and indeed the arguments made in the existing literature, including those by Reviewers 1 and 2 – does not rest on causality. That is, while there may have been some confusion about whether we were making claims of causality, our core argument that AMP effects are primarily associated with awareness of the primes is agnostic to causality. Now that we have corrected our language around implications of causality, we hope that the reviewers can more fully engage with our core claims and evidence that participants are aware of the influence of the primes on their evaluations of the target stimuli.

We believe that we demonstrate this across our original five experiments, while also taking take the point that Reviewer 2 noted: that these experiments do not rule out the possibility of post-hoc confabulation. However, our new 6th and 7th Experiments now directly address and remove this possibility (particularly our 7th Experiment, where confabulation is simply not possible). We believe that the implicit misattribution account cannot accommodate for these findings as readily as the explicit account. As such, we believe these findings now represent a significant weight of evidence within the debate around the automaticity of AMP effects and therefore the implicitness of the task.

**Editor**: Finally, I want to add a note about what I would expect in a revision. As noted by both reviewers, many of the issues raised here have already been hashed out in the previous back-and-forths about the AMP. And as I wrote, the data do not warrant sweeping conclusions such as the title's provocative implication that the village of people who use the AMP have collectively decided to ignore its naked absurdity.

**Authors**: While we agree that these issues have seen debate in previous papers, we strongly disagree with the idea that the debate is therefore settled. Indeed, one important features of our papers is to highlight important methodological, statistical, and conceptual issues which are prevalent in this literature which make it difficult to come to meaningful conclusions. This issues bring the conclusions previous back-and-forths as a whole into question given that, in many cases, they rest on in appropriate methodological or statistical conclusions.

In terms of the title, we regret that readers have taken any offense. In our previous manuscript, we were explicit in the discussion section that the analogy with the parable of the Emperor’s New Clothes is in the risk of future work not taking new evidence on board, rather than commenting on any previous work. And, to put our own cards on the table, we are suckers for puns and it made us laugh. We hear that others have not had the same reaction and very much take the feedback. Our revised title is “The The AMPeror’s New Clothes: Performance on the Affect Misattribution Procedure is Mainly Driven by Awareness of Influence of the Primes

**Editor**: But, some valid questions do remain about possible limitations of the task, when and how awareness matters (or not), and some detailed mapping of boundary conditions regarding when and for whom misattribution is more or less likely to occur. I think a careful, measured interrogation of these issues would be quite valuable to researchers and make a solid contribution to the literature. I would welcome it at JPSP-ASC if you believe you can make such a contribution.

**Authors**: We agree that such questions are highly relevant to the literature. We believe now, in light of our new experiments ruling out post-hoc confabulation as an explanation for our observations,

We speak directly to the issue of possible limitations of the task (i.e., it does not reflect automatic responding the way it has previously been claimed, and because of this likely does not reflect misattribution as is traditionally conceived).

We also speak directly to how awareness matters (because the AMP only meets criteria for structural validity when responding is predominantly influence-aware, and the fact that most AMP effects/the AMP’s predictive utility appears to be based on influence-aware trials suggests AMP effects do not reflect automatic evaluations).

We do not delve into the mapping of boundary conditions of misattribution, because ultimately we are not interested in misattribution: we are primarily interested in demonstrating that AMP effects are not produced under an automaticity condition which they are regularly assumed to be produced under.

This has implications for the misattribution mechanism, but ultimately whether misattribution occurs, and to whom it occurs for, is outside of the scope of this manuscript. However, we include a detailed discussion on the idea that future work should assess whether influence-awareness is a trait-like individual difference (as it appears to be). This line of research is one which we are also currently pursuing, but ultimately we believe it falls outside of the scope of the current manuscript.

**Editor**: Overall, then, though the current version of the paper is not acceptable, I encourage you to pursue this line of work by following the detailed and thoughtful suggestions made by the reviewers. I will keep the door open to a new submission if you can address all issues with additional data and substantial re-organization of the paper. If you choose to take this route and have completed such a revision, please submit the revision as a new manuscript. This submission must be accompanied by a detailed cover letter indicating the prior history of the paper. You should also explain which specific changes you made and which recommendations you did not follow and why. This letter should address all of the points raised in my decision letter plus any other major, non-redundant points mentioned by each reviewer.

In closing, thank you for submitting to JPSP-ASC. I would also like to thank the reviewers for their service to the field. Their thoughtful comments and suggestions were very helpful in reaching my decision.

Sincerely,

Elliot T. Berkman

Associate Editor

Journal of Personality and Social Psychology: Attitudes and Social Cognition

Reviewer #1: Signed: Yoav Bar-Anan

The manuscript reports five experiments in which the authors tested the relation between the priming effect in the AMP and awareness of the priming effect in the AMP. In Experiment 1, the authors introduced a modification of the AMP, in which, in 120 trials, after evaluating the target Chinese pictograph as Pleasant or Unpleasant, participants were requested to "Press spacebar if the picture influenced your response to the Chinese symbol" within a 2000ms after they evaluated the Chinese pictograph. The primes were IAPS. After the AMP, participants reported on 1-7 scale, the extent to which the primes influenced their ratings of the targets. Priming was stronger on trials in which participants pressed space than on the other trials. Participants who pressed the space more often, showed stronger priming. In a multiple regression, the rate of pressing space and the retrospective response to the influence question at the of the experiment (these measures had a correlation of .78), both predicted the size of the priming effect.

In Experiment 2, before the modified AMP (which was the same but with 72 trials), participants completed an IAPS AMP that did not include the awareness check after each trial. The authors found that the rate of reporting the influence in the modified AMP predicted the size of the priming effect in the previous AMP. The authors also reported that the priming effect in the non-modified AMP were stronger than the priming effect in the modified AMP, computed only from trials in which participants did not report influence of the primes.

Experiment 3 was identical to Experiment 2, but instead of IAPS, the first (non-modified) AMP had photos of Obama and Trump as primes, and all the participants identified as supporters of the Democratic party. The results were the same as in Experiment 2.

Experiment 4 was similar to Experiment 3, but the two AMPs were the modified AMPs, and participants reported their political preference. The authors reported that they replicated the results of the previous experiments. They also found a correlation between the rates of reporting awareness after the trials of each AMP (r = .82), and the rate of influence reporting in each AMP predicted the size of the priming effect in the other AMP. The authors also found that trials about which the participant reported a priming effect were better at discriminating between self-reported support of the Republican party versus the Democratic party (d = 2.08) than trials about which the participant did not report priming (d = 0.62).

Experiment 5 was similar to Experiment 2, but the AMPs included the modifications recently recommended by Mann et al. (2019): There were only 60 trials in each AMP, the instructions emphasized more strongly than usual that participants should not rate the primes, and the targets were paintings rather than Chinese pictographs. The results were the same as in Experiment 2.

In a meta-analysis of the five experiments, the authors reported that 54% of participants reported priming in 0-20% of trials, 14% reported priming on 21-40% of trials, 8% reported priming on 41-60% of trials, 6% reported priming on 61-80% of trials, and 17% reported priming on 81-100% of trials.

In another meta-analysis of the five experiments, the authors reported that the average rating of the targets after positive primes and the average rating of targets after negative primes were positively correlated when computing those averages only from trials in which participants did not report a priming effect, and were negatively correlated in trials in which participants reported a priming effect.

The authors concluded that the priming effect in the AMP is not implicit: "there is no clear evidence for [the priming effect] being unintentional, and new evidence against being unaware".

1. The manuscript has a great potential to make a positive contribution to the scientific community. The main strength of the manuscript is the finding that reporting the priming effect in one AMP predicts the priming effect in a previous AMP. Other informative findings are the conceptual replication of the positive relation between the priming effect in the AMP and retrospectively reported priming, and some evidence that might suggest that good psychometric qualities in the AMP depend on a minority of the participants - those who report the priming effect. The authors also provide an interesting discussion of previous results and interesting strong opinion on how these findings should influence researchers who use the AMP.

This manuscript is a clear challenge of the validity of the AMP, and researchers should be exposed to that challenge, to help them decide whether to use the AMP, and how to interpret results obtained with the AMP. Personally, my conclusion about the AMP has not changed: it is one of the best indirect measures of evaluation we have, but that's only because we do not have good measures. Like the other implicit measures, its validity is highly questionable, and inference from results obtained with the AMP is currently very tentative. I agree with the authors that many publications do not seem to exercise the appropriate caution when interpreting AMP results, and I believe that this manuscript could help raise awareness about the possible weaknesses of the AMP. It will be highly cited and could have a very positive impact on people's understanding of the AMP.

**Authors**: We thank Reviewer 1 for his kind words and assessment of our paper.

**Reviewer 1**: Notwithstanding the great potential of this manuscript, it has some weaknesses that might damage the readers' understanding of current evidence about the AMP. In the rest of this review, I will list a few comments and suggestions that the authors might consider in a possible revision, all with the purpose of improving the service the manuscript would provide to the readers, and minimizing possible negative effects.

The authors argue that retrospective awareness of the priming effect suggests that misattribution does not underlie the mechanism. They argue that misattribution requires unawareness. This seems logical: if one is aware of a misattribution, then one can correct that misattribution before responding. However, this is not definite. First, awareness might have risen only after observing the response. In fact, awareness might not occur at all unless prompted with the direct question about the priming effect.

**Authors**: Reviewer 1 first asks if awareness only arises after observing one’s response. To test this idea we carried out two new studies (Experiments 6-7). Here we further modified our IA-AMP so that participants had to indicate if they were aware of the prime’s influence on their target evaluation either (a) before the target evaluative response was emitted (Experiment 6) or (b) before the target stimulus was even presented onscreen (Experiment 7). In these studies awareness occurs prior to the response as the response (and in the latter experiment target stimulus) has yet to even occur. Even under these conditions we still obtain similar effects as reported in Experiments 1-5.

Reviewer 1 also asked whether awareness might not occur at all unless promoted with the direct question about the priming effect. If awareness was simply an artifact of our IA-AMP and unrelated to standard AMP effects then it should not have backward predicted standard AMP effects where awareness was never probed. Yet that is precisely what occurred in XX of our seven studies. Thus we think it unlikely that awareness was simply a post-hoc irrelevance and more a core factor moderating AMP effects.

**Reviewer 1**: Second, participants could suspect that the prime influenced their evaluation of the target even before they rate the target, but without any choice other than evaluating the target, there is little reason for them to reverse their response (e.g., from Pleasant to Unpleasant). In other words, in the AMP, participants cannot avoid misattributing even if they suspect that it occurred. More broadly, being able to detect misattribution does not mean that people know how to correct for it. Thus, I am not sure that what the authors present as the most likely conclusion from their findings (misattribution does not underlie the priming effect) is the only possible conclusion. It is definitely a plausible conclusion - plausible enough to cast serious doubt on the AMP's validity, but readers would benefit from exposure to other possible conclusions.

**Authors**: On the one hand, we agree with Reviewer 1 that misattribution could play a role in IA-AMP effects if one makes a number of post-hoc adjustments to the concept and how it is traditionally conceived. Specifically, that valence is misattributed from the prime to the target and that even though people know this is happening they still do so – because they feel compelled to do so, they have no other information to go on, feel that this is the experimental goal and they want to be good participants, or a host of other reasons (for more on this see the “Do AMP effects reflect a misattribution process?” section in our General Discussion).

On the other hand, however, all of these possibilities are post-hoc justifications for our findings and nor *a priori* claims made in previous empirical papers or theoretical ones. As far as we are aware, past work on the AMP effect and misattribution as a mental process do not claim that misattribution occurs intentionally and in the presence of awareness (CITE SPECIFIC PAPERS HERE). Therefore, while we are happy to entertain post-hoc amendments to a theoretical concept, those post-hoc justifications should not be treated as equivalent to *a priori* pre-registered claims that conflict with them.

Instead, it should be acknowledge that an existing theoretical concept does not – as it stands – predict nor explain our findings. Instead, that a post-hoc adjustment to that concept could modify it in such a way that a plausible explanation could be offered. But any such adjustment (i.e., that misattribution occurs with awareness and intentionally) deviates significantly from what went before and should be empirically investigated rather than solely conjectured.

**Reviewer 1**: Still on the same subject, in the modified AMP, participants could use the compatibility between the valence of the prime and the valence of their rating as evidence for the influence of the prime on the target. Therefore, even if participants have no awareness of the priming when it occurs, they could still respond based on that compatibility. Further, it seems reasonable that people would detect a compatibility between their rating of the target and the valence of the prime more frequently when participants are more sensitive to the AMP (e.g., to misattribution). In other words, if some participants are more likely than others to show priming in any AMP, they would also be more likely to report the priming (in any AMP). Therefore, the finding of a positive relation between the awareness in the modified AMP and the priming effect in another non-modified AMP is not unequivocal evidence that misattribution is not responsible for the priming effect in the AMP. Again, the authors' account is plausible and important to share because it has serious implications, but the readers would also benefit from an explicit reminder of alternative accounts.

**Authors**: Our sixth and seventh experiment directly speaks to (and contradicts) this alternative interpretation. However, the introduction of the section for the sixth experiment provides a reminder to readers that alternative perspectives could be possible.

**Reviewer 1**: Related to the previous point, in p. 15, the authors wrote that they sought to determine if awareness drives AMP effects. They then use the verb "driven" often throughout the manuscript. I think that "drive" implies a causal role for awareness. However, the authors did not manipulate awareness. Therefore, they can conclude only about the possibility of a relation between awareness and the AMP effect, and not a causal relation. Very often, the word "drive" seemed inaccurate and might have conveyed the wrong message. Often, moderation of the priming effect by reported priming was described as evidence that the priming effect was driven by awareness or by trials in which participants showed awareness, or by participants who reported much awareness. It is possible that I do not understand the meaning of "drive", but I do not think that it is common to describe findings of moderation, especially when the moderator is not manipulated, as evidence that the moderated effect is driven by the moderator.

**Authors**: We take the reviewer’s point that our use of “driven” was at times misplaced. We still hold firm that the AMP effect (at the group level) is driven by the subset of participants who are more frequently aware of the influence of the prime on the responses (because this claim is at the effect level, and is demonstrated across our studies).

Nevertheless, and in line with Reviewer 1’s suggestions, we have revised our language throughout the paper (i.e., STATE EXACTLY HOW HERE).

**Reviewer 1**. The authors seem to accept the idea that in order to measure implicit cognitions (e.g., attitudes that influence behavior without people's awareness), the mechanism that underlies performance in the measure must be implicit (e.g., the priming effect in the AMP must occur without people's awareness). Clearly, this is not always the case for psychological measures. When I report that I strongly agree with the statement "I am shy" in a shyness questionnaire, it is likely that none of the processes that cause my shy behavior also cause my response in the questionnaire. This might be also true for the IAT and evaluative priming: it is possible that the processes the mediate the effect of mental associations on performance in those tasks are quite different from the processes that mediate the effect of mental associations on automatic evaluation.

The authors might argue that if the priming effect in the AMP elicits awareness, there is little reason to suspect that the AMP would measure evaluation that escapes awareness. That might be so, but, by now, there is published evidence about the validity of the AMP as a measure of automatic evaluation that go beyond the investigation of the processes that underlie the priming effect in the AMP (for reviews, see Cameron,Brown-Iannuzzi, & Payne, 2012; Payne & Lundberg, 2014 [see the validity section]). It would benefit the readers if the authors acknowledge that. The authors could also choose to review that evidence and cast doubts on their validity (e.g., I have not seen any convincing finding that was replicated in an independent lab). Yet, at this time, even a finding that the priming effect in the AMP is completely intentional would not suffice for the conclusion that it is not a good measure of automatic evaluation, without arguments against the evidence reported so far from (mostly correlative) validation studies that helped establish the AMP as a measure of implicit social cognition.

**Authors**: We would contend that much (if not all) of the published literature which has provided evidence for the AMP’s validity as a measure of automatic evaluations has suggested that the AMP is valid based on the fact that (i) it predicts criterion behaviours, and (ii) its scores converge and diverge between and from explicit measures under theoretically-predictable conditions.

The first of these two criteria are based on external validity; i.e., the AMP’s relationship with other measures external to the AMP itself. However, the fact that the AMP predicts criterion behaviors is not evidence for its validity as an *automatic* measure of cognition. Additionally, based on the second point, although convergence with/divergence from explicit measures in theoretically-predictable ways is stronger evidence, other research using the AMP has demonstrated that implicit-explicit correlations between measurement procedures (or lack thereof) are not necessarily indicative of correlations between the relevant latent attitude constructs. This is because of the fact that these correlations can often be traced back to the structural (dis)similarity of the measurement procedures (Payne, Burkley, & Stokes, 2008). As well as this, such correlations are often attenuated due to measurement error associated with the measurement procedures (Archer et al., 2008). For these reasons, we feel that these points cannot necessarily be taken as strong evidence for the validity of the AMP as a measure of automatic evaluations. There is also other published evidence that does not find the AMP functions well in predicting automatic evaluations (e.g., Teige-Mocigemba et al., 2018).

In our structural validity analysis, we investigate whether the AMP is a valid measure of *automatic* evaluations (in the sense of the evaluations of the stimuli within the procedure occurring without awareness). We demonstrate that the AMP *is* a valid measure of evaluations when scores are calculated using non-aware trials, but the AMP *is not* a valid measure of evaluations when scores are calculated using aware trials. So the AMP can be either a structurally-valid measure of nonautomatic evaluations, or a structurally-invalid measure of automatic evaluations. This can explain why the AMP shows correspondence with criterion behaviors (i.e., it is validly measuring evaluations under nonautomatic conditions), and also why the AMP is only effective for subsets of individuals (i.e., it is more valid for those participants who are influence-aware).

**Reviewer 1**: The description of Experiment 2 in Payne et al. (2013) does not seem accurate. To the best of my understanding, the most important finding was that the AMP predicted judgment of a Black (but not White) target that behaved ambiguously, whereas the direct rating of the primes did not. I think that this is one of the best findings in support of the AMP as a measure of an implicit construct (and pursuing its replication should be a priority of our field, especially considering the rather small sample in the original experiment, n = 45). In the first description of this experiment in the present manuscript (pp. 8-9), that aspect of the experiment is not mentioned at all.

**Authors**: We thank Reviewer 1 for pointing this omission out to us. Upon reading the Editor and Reviewers 1-2’s comments we realized that when providing details of issues in prior studies we failed to provide an overview of *other* effects observed in those same studies which were not flawed. We have revised our manuscript to in order to provide more comprehensive detail of past work (see revised material on p.XX).

**Reviewer 1:** Later (pp. 13-14), the authors wrote that Payne et al. "based their inference on the fact that there was a significant difference between personality judgments and 'intentional' AMP effects, but no significant difference between personality judgments and 'unintentional' AMP effects". But it is unclear what they mean by "difference". The test in question was of a relation between the AMP effects and the personality judgment, not of a difference between them (it would also be unclear to the readers what the authors mean by "personality judgments" because this aspect in the experiment is never described in the present manuscript).

**Authors**: In their paper, Payne et al. wrote: “Second, direct (intentional) ratings were more affected than indirect ratings by previously rating the personality of a Black person in comparison with a White person”. As we understand it, this inference is derived from the following analysis in their Results section: “As expected, direct test scores were lower (indicating less stereotypical judgments) in the Black impression character condition (M = 0.02, SD = 0.43) than the White character condition (M = 0.46, SD = 0.53), F(1, 43) = 9.51, p < .01. No such difference was found for the indirect test scores, F(1, 43) = 0.92, p = .34.”.

If we are correct, and this conclusion was derived from the statistics reported above, then the authors cannot infer that direct ratings were *more affected* than indirect ratings, because the authors do not directly compare the two measures. In order to do so, the authors might have (for example) conducted an ANOVA with ratings as a DV, character condition as one IV, and rating type (direct vs. indirect) as another IV. If the authors had found an interaction effect between character condition and rating type, then this may have allowed for the above inference to be made (assuming effects were in the relevant predicted direction). However, the authors did not make such a comparison. Therefore, we stand over the believe that their inference here was invalid.

**Reviewer 1**: The description of the results and conclusions of Experiment 3 in Payne et al. (2013) do not seem accurate. The authors wrote: "Even though there was no way to determine what proportion of AMP effects were driven by aware vs. non-aware trials (given the necessary data was not collected), the authors still argued that effects on the traditional AMP did not differ from those on the modified AMP, and used this as evidence for the relative unawareness of the AMP."

First, the comparison between the AMP with and without the option to skip trials in which the participant suspect a priming effect is informative. Had Payne et al. (2013) found a reduction in the priming effect in the modified AMP, in comparison to the traditional AMP, that would have supported (to some extent) the argument the priming effect in the AMP requires awareness.

**Authors**: We agree with Reviewer 1 (and indeed, Payne et al.) that, had such a reduction been found, this would have provided evidence that “the priming effect in the AMP requires awareness”. It was this very logic from which we started and designed the studies reported in Experiments 1-7. Thus if the skip trials provide evidence for prime influence awareness in the Payne et al. study we see no reason why the nearly identical procedure used in our studies should be viewed any differently. In other words, the only difference between the procedure used here and that of Payne et al. is that instead of skip the trial and not record the evaluative response, we skipped the trial but recorded the response. In this way we obtained not only the skip data but the evaluative response as well. In short, if one views the procedure the Payne et al. used as sensitive to prime influence awareness, as Payne et al. and Reviewer 1 seem to do so, then we see no reason why our procedure (and by implication our findings) should be viewed any differently given that it involves a near identical procedure.

**Reviewer 1**: Surely, under NHST, lack of significant evidence is less definitive than finding significant evidence, but that is not related to the lack of appropriate comparison (further, Payne et al. addressed the issue of statistical power in their discussion of the results of that experiment, p. 383).

Payne et al. made the following inference based on their results. They claimed that “th[e] opportunity for selective responding did not eliminate *or* *even reduce* the priming effects”. As far as we can tell, the italicised claim is not substantiated by the analyses which were conducted by the authors in their paper. This is because the authors never directly compared the magnitude of the two effect sizes. Rather, the analyses which the authors did report merely demonstrated the presence of an AMP effect for both AMP versions. Thus, the first part of their claim (“did not eliminate the priming effects”) is valid but the second (“or even reduce”) is not. If the authors wished to substantiate this claim they need to provide (i) a direct comparison of the effect sizes (ii) using Bayesian analyses (or Frequentist equivalence tests) to provide evidence for the null hypothesis. They did not do so. Thus the first part of their inference is valid whereas the second is invalid. This is not to say it could not be tested given their data – simply that they did not appear to do so in their paper.

We have revised the manuscript to better clarify this point (see revised on pp.XX)

Payne and colleagues did indeed address the issue of statistical power in their discussion.

**Reviewer 1**: Second, and perhaps more important, the authors ignore a major finding in Payne et al.'s (2013) Experiment 3: "Participants passed much less when the primes were pleasant (M = 0.14) or unpleasant (M = 0.17) than when the prime was neutral (M = 0.54), F(2, 70) = 28.23, p < .001. Passing rates on neutral trials were significantly higher than pleasant trials, F(1, 35) = 34.0, p < .001, or unpleasant trials, F(1, 35) = 25.65, p < 001". Clearly, that pattern is the opposite of real awareness of the priming effect. Why would there be more priming when the prime was neutral rather than of clear valence? Payne et al. (2013) proposed a plausible explanation: when priming occurs, participants feel (because of misattribution) that they have clear evaluation of the target. When priming does not occur, participants are less convinced regarding their evaluation of the target, and are more concerned that the prime influenced that evaluation.

**Authors**: We acknowledge that we failed to report this finding in our original submission. We now do so in the revised manuscript (see revisions on pp.XX).

The reviewer asks “Why would there be more priming when the prime was neutral rather than of clear valence?”. The authors indeed proposed one plausible explanation for this, but there are also others. To take a step back briefly, the implicit misattribution account assumes that AMP effects arise from participants attempting to follow AMP instructions, which should not result in any prime-consistent effects, but prime-consistent effects occur anyway due to the unintentional and unaware influence of the primes on participants’ evaluations of the targets. This account can provide an explanation for the increased frequency of skipping on neutral trials relative to valenced trials.

The explicit account of AMP effects assumes that AMP effects arise because a subset of participants, on a subset of trials, intentionally and with awareness use the prime’s valence to inform how they respond to the target stimuli. In this case, participants respond counter to the instructions of the AMP: in spite of being told not to use the prime’s valence, they do anyway. In this case, the explicit account also can provide an explanation of the increased frequency of skipping on neutral trials relative to valenced trials. Specifically, if the explicit account is correct, then it would also make sense that they would skip more on neutral trials because they have no evaluative information to use in responding to the targets. Of course, this assumes participants use the skip trials differently to how they are instructed to be used. Yet this is already happening according to the explicit account. So the propensity for participants to skip more for neutral primes than valenced primes can be explained easily by either account. The evidence is not particularly persuasive in either direction, in our view.

We have elaborated on this issue in the conceptual issues section of our introduction (see revisions on pp.XX).

**Reviewer 1**: To conclude points 5 and 6, the weaknesses the authors found in Payne et al.'s (2013) research are not very convincing, and also seem to rely on inaccurate or incomplete description of Payne et al.'s studies. As a slight digression, I would add that this flaw in the present manuscript is unfortunate because Payne et al.'s (2013) studies had several weaknesses. In Experiment 1, the fact that some participants reported unintentional rating of the primes does not preclude the possibility that other participants rated the primes intentionally (i.e., perhaps those who report intentional and those who report unintentional priming are not the same people). For Experiment 2, if the priming effect is driven mostly by a minority of participants who choose to intentionally rate the primes, then the AMP is not exactly the same measure as a direct rating of the primes. For instance, perhaps, unlike direct rating, most of the variance in the AMP comes from people who do not try to hide their preference for one social group over the other. That difference between the AMP and direct rating of the primes could be the reason why the AMP is sometimes better than direct rating in predicting race-related behavior. For Experiment 3, if the priming effect is driven mostly by a minority of participants who choose to intentionally rate the primes, then it seem likely that these people would not want to use the option to pass trials in which the primes influence their rating of the targets. As a result, that modification of the AMP would not be effective in eliminating intentional rating of the primes.

**Authors**: In line with Reviewer 1’s suggestions we have revised our paper to more accurately characterize what others have found and reported. At the same time, we want to acknowledge that there still remains statistical, conceptual, and methodological issues with that work, that call into question the strength of the claims being made. This held true before and still, as far as we can see, does so.

Additionally, we have added some of the reviewer’s points on the limitations of these experiments within our introduction (with reference to the fact that it was the reviewer who suggested them).

**Reviewer 1**: It was not entirely clear what methodological shortcomings Gawronski & Ye's (2015) research had. Their crucial finding was that the retrospective reports of the priming effect correlated with the priming effect only for the topic that was salient during the task, and not for the topic that was not salient. If the reason for the correlation between the priming effect and retrospective reports of the priming is due to intentional rating of the primes, why would the manipulation of topic salience influence this correlation without influencing the priming effect itself? The present authors wrote "retrospective self-reports do not provide a direct assessment of the construct under investigation". Yet, Gawronski and Ye did not rely on those self-reports as a measure of awareness of the priming effect. Rather, they tested whether the finding of a correlation between retrospective self-report and the priming effect survives a certain manipulation of awareness. They showed that their manipulation of awareness decreased the validity of the self-reported awareness of the priming effect as a predictor of the priming effect but did not decrease the priming effect itself (the results summarized in Table 1 in Gawronski & Ye's article are the best evidence I have seen so far, against the intentional rating account). It seems reasonable to conclude from that evidence that the self-reported awareness of the priming was not due to a necessity of awareness for the priming effect to occur.

**Authors**: We assume that the reviewer means “intentionality” instead of awareness within his comment here (since Gawronski and Ye were interested in the former, not the latter).

We noted a number of issues with the Gawronski and Ye paper – some conceptual and other statistical. We consider both in turn.

**Conceptual issues**. Our point is that, because post-hoc self-reports asking for reporting on *global* performance within a task are questionable, change in these scores on the basis of an (assumed) manipulation of intentionality cannot be used to infer that intentionality changed, as it is unclear what this scale actually measures.

Additionally, the authors assumed that their manipulation of intentionality functioned well as such. However, the authors rely merely on their speculation that intentionality *should* be manipulated by their interventions. But they never directly test this claim. Given recent work demonstrating that such bespoke manipulations are often poorly-validated (see Chester & Lasko, 2019), the conclusions that can be drawn from this work are limited at best.

**Statistical issues**. The authors also committed the inferential fallacy of concluding a significant difference from a nonsignificant (“marginal”) test statistic.

In light of Reviewer 1’s comments we have revised the manuscript to better clarify these various points (see changes on pp.XX).

**Reviewer 1**. The authors conclude that the AMP priming effect "do not represent an equally valid measure of attitudes across individuals". This seems a valid conclusion from the evidence they report, and it is compatible with the evidence reported in Bar-Anan & Nosek (2012, 2014). In our 2012 research (mainly in Tables 3 and 4), we showed that indices of psychometric quality are reduced when excluding from the analyses participants who reported intentional rating of the primes (or, at least, awareness of the priming effect). We also found (see Appendix D of Bar-Anan & Nosek, 2014, Figures A and B, at https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-013-0410-6/MediaObjects/13428\_2013\_410\_MOESM1\_ESM.pdf) that the AMP loses its relation with direct measures of evaluation much faster than other indirect measures, after removing participants with extreme scores (those with the largest priming effects). However, all that evidence is still insufficient to inform us how serious this problem is. Only the appendix from our 2014 paper provides some comparison with other indirect measures (and the AMP seems inferior to the other measures). Yet, I did not see much research about how many participants "drive" typical effects in social psychology, and how many are the main contributors to validity evidence of psychological measures. I also do not know of much research that informs us how inequality in validity of a measure across individuals affects scientific progress. Clearly, it is better if a measure works well for a larger portion of the population, but what is the standard and how much does scientific progress suffer from each drop in that equality? I think that readers would need that knowledge in order to make strong conclusions about the implications of the inequality reported in the present manuscript.

**Authors**: We are glad that Reviewer 1 agrees with us that our evidence highlights that the AMP effect is not an equally valid measure across all individuals.

First, AMP effects are commonly-used to make inferences about *automatic evaluation* in the general population (CITATIONS!). If those effects do not function as a valid measure of people in general then population level claims are invalid*.*

Likewise, AMP effects are commonly used to make inferences about mental mechanisms operating in people in general. Once again, if our analyses are correct, then such claims are invalid. In other words, if AMP effects reflect the evaluations only a subset of individuals (approx. 20% of participants) then they cannot be used to make inferences about mental processes in people *in general*.

Additionally, in those who do not show the effect, responding may be automatic, but it does not represent evaluations (since it fails to meet a key criterion of its structural validity). For participants who show an effect, responding is evaluative, but not automatic.

In short, “inequality in the validity of the AMP across individuals” clearly represents a significant barrier to scientific progress that wishes to make claims about automatic evaluation and mental mechanisms (which the vast majority of AMP studies do). For a given participant, we are either studying evaluation or automaticity using the AMP – but likely never both.

We have revised the manuscript to better acknowledge these various points (see changes on pp.XX).

Finally,– how prevalent are such subset effects in other procedures within psychological research? However, this question is beyond the scope of this work. We would also suggest that the “standard” does not matter here. Validity is, ultimately, not relativistic: a measure cannot be considered valid just because it is “less invalid” than other measures.

**Reviewer 1**: In the "Structural Validity" section, the authors seem to expect a negative correlation between rating of targets after positive primes and rating of targets after negative primes. That would be the case mostly if priming is the main factor that influences the rating of the targets. However, there might be other factors that influence the rating of the targets. If that is the case, then controlling for those factors would be useful for a better measurement of the construct reflected by the priming effect. By comparing two categories of prime stimuli (e.g., positive and negative primes), one can minimize the effect of non-evaluative factors that influence the rating of the targets (e.g., liking of the Chinese culture, and a general tendency to rate stimuli as positive or negative). In other words, the measure of evaluation in the AMP is not the average rating of the targets after a certain category of primes. It is the comparison between the average ratings of the targets after one category of primes and the average ratings of the targets after another category of primes.

For that reason, I did not accept the authors conclusion that "while it could be argued that non-influence aware trials on the IA-AMP represent 'implicit' responding, these trials do not function as a structurally valid measure of evaluations. " (p. 53).

**Authors**:

**Reviewer 1**: Somewhat related, I do not think that the authors were accurate when they wrote that "the primes only exert influence on ratings within the AMP task when participants are highly influence-aware." Figures 2 and 3 suggest that priming occurred even when participants report no awareness of the priming effect. Further, although throughout the manuscript the authors often did not report the priming effect in "unaware" trials, whenever they reported that effect, it was significantly larger than zero (in p. 29, the effect was d = 0.82; in p. 38, the effect was d = 0.62).

For a similar reason, I think that the authors are inaccurate to conclude, in p. 56, that for the majority of participants, scores cannot be said to represent a sound measure of evaluations at all. Unless I am missing something, Figure 3 seems to suggest that most participants show the priming effect, which reflects evaluation.

Authors: The claim Reviewer 1 mentions (i.e., that *the primes* only exert influence when participants are highly influence-aware) is, in our eyes, accurate.

We are not arguing that no effect is present when participants are not highly influence-aware. Instead we are arguing that the presence of *any* effects in influence-unaware participants cannot be attributed to the primes themselves, because of the fact that the ratings of the primes are not negatively correlated (and thus, the primes themselves do not appear to be exerting influence on responding).

We should also note that the analyses Reviewer 1 highlights for the presence of AMP effects refers to analyses carried out at the *trial-level of analysis* whereas the discussion here is focused at the *participant-level of analysis* (where effects are related but distinct).

In short, we are not arguing that AMP effects are evident only in those who are highly influence aware. Rather we are arguing that for those who are influence-unaware, the AMP effect does *not* reflect evaluation of the primes (because the measure does not demonstrate an aspect of structural validity it has been (implicitly) assumed to meet).

**Reviewer 1**: In p. 21, when the AMP is first described in the method, I recommend providing more information about the procedure (trial sequence, block sequence, and procedure sequence) rather than refer the readers to a different paper.

**Authors**: This has now been added (see changes on pp.XX)

**Reviewer 1**: In p. 21, I was confused by the authors' description of the most crucial modification of the AMP: "rather than allow participants to skip trials if they felt that they would be influenced by a prime, we instead asked them to respond to every trial (i.e., "Press spacebar if the picture influenced your response to the Chinese symbol"), and thereafter indicate if that response was influenced by the prime (i.e., by pressing the spacebar during a fixed 2000ms post-response interval)." It seems that the instruction that appear to describe the request to respond to every trial is the instruction relevant to the awareness question. I had to read the Inquisit script (provided in online materials) to make sure I understood the task correctly.

**Authors**: The manuscript has been revised to clarify this point (see changes on pp.XX).

**Reviewer 1**: It would probably be helpful to most readers, if the authors provide clearer descriptive statistics for all their studies. In each experiment (and not only meta-analytically), I was particularly interested in the mean and SD priming effect for "unaware" and "aware" trials (and perhaps more details about the full distribution), the mean and SD number of "aware" trials, and a scatter-plot showing the relation between the percentage of "aware" trials and the priming effect in the same IA-AMP, and in the other AMP (Experiments 2-5). With those descriptive statistics, readers would have a much better understanding of the findings, beyond the results of the statistical tests.

**Authors**: These have now been added.

**Reviewer 1**: I applaud the authors for pre-registering their experiment and providing full access to their materials, data, and analysis. It is important to publish papers that follow these new norms. However, I was unable to find clear reports of the analyses that, according to the authors, were supposed to appear in the Supplementary Materials on OSF (e.g., footnote 8, a few times in p. 32, and once in p. 36). Perhaps the authors mean that these results appear in the html file produced by RStudio from the analysis scripts. I think that it would be better to provide a clear document (Word or PDF) with a summary of all the additional statistical analyses.

**Authors**: Reviewer 1 is correct that the Supplementary Materials refer to the html Markdown files produced by the analysis files. While we prefer to use these html files (rather than Word or pdf files) in order to ensure that these reports are also reproducible, we have added a footnote in the manuscript to clarify what Supplementary Materials refers to (see changes on pp.XX). Note that outputting Word or PDF versions of these files is possible through the use of RMarkdown within our original analysis files.

**Reviewer 1**: In p. 45, the authors report the trial-level meta-analysis but refer the readers to Figure 2, which seems to show participant-level results.

**Authors**: The paper has been revised as requested (see changes on pp.XX)

**Reviewer 1**: In p. 45, to interpret the moderation of the priming effect in each trial, by the self-reported awareness of the priming effect, the authors compared the moderation effect-size and the priming effect-size. That is interesting, but, usually, moderation is explained by reporting the simple effects in different levels of the moderator. In this case, it seems essential to report the priming effect in trials that ended with a space response (i.e., self-reported priming) and the priming effect in trials that ended without a space response (i.e., trials in which the participant did not report an influence of the prime on the rating of the target).

**Authors**: We have now added the mean evaluation of each prime type at each level of influence-awareness (see changes on pp.XX).

**Reviewer 1**: P. 11: "Dietvorst and Simonsohn (2018) recently found that people readily incorporate to-be-ignored information into their responses on different tasks, despite the fact that researchers signal that this information was irrelevant and to be ignored". Does "readily" mean "intentionally"? If it occurs unintentionally, then this finding does not provide support for the authors' suspicion that participants ignore the instructions in the AMP, and intentionally use their evaluation of the primes when they rate the targets.

**Authors**: “Readily” does means intentionally (and the authors of the referencing paper explicitly investigated the intentional usage of to-be-ignored information. We have revised this sentence to specifically include the use of “intentionally” to avoid any potential ambiguities in what we are referring to (see changes on pp.XX).

**Reviewer 1**: In p. 28, the authors reported "Consistent with Experiment 1, we found that IA-AMP effects were driven by that subset of trials where participants reported being influence-aware, OR = 20.65, 95% CI [17.10, 24.94], p <.001, Cohen's d = 1.67, 95% CI [1.57, 1.77]." I assume they meant that reporting awareness of the influence of the primes moderated the effect of the prime valence on the target evaluation. This is not clear, currently. And, as noted earlier, moderation is not evidence that an effect is driven by the moderator. It is only evidence that the moderator moderates the effect.

**Authors**: We have now revised the paper as requested (see changes on pp.XX).

**Reviewer 1**: In p. 36, participants chose not to report in the main manuscript the results that replicated the relation between reporting priming and the priming effect (on the trial-level and on the participant-level). These results seem rather central to the present manuscript, so I suggest including them in the main text (if the results are complex or seem repetitive, a table might help).

**Authors**: We now include the descriptive statistics for this replication analysis in each experiment (see Tables XX-XX on pp.XX). We do not report the *p* values for the results of these analyses (they are available in the supplementary materials). We hope the inclusion of these descriptive statistics, coupled with the statement that the effects replicated, and the meta-analytic effects in the meta-analysis section, will satisfy the reviewer.

**Reviewer 1**: Experiment 4 provides an opportunity to examine whether reported priming equally predicts the priming effect in a subsequent and in a preceding AMP. In other words, it might be informative if the authors add the order of the tasks as a factor (and a moderating factor) in the multiple regressions reported in pp. 36-37. That would further test the bidirectionality of the relation between reported priming in one task and the priming effect in another task.

**Authors**:

**Reviewer 1**: In Figure 1, the labels were not immediately clear to me. The x-axis showed the priming effect, reflecting preference for Trump over Obama. The graph included labels to explain the meaning of the two most extreme possible scores (-1 and 1). However, those labels were not perfectly clear, and it was not clear that these labels were supposed to reflect the values -1 and 1. Instead of using those labels, it is common to simply explain, in the Figure's note, what a positive score reflects.

**Authors**: We have revised the description of the figure to explicitly describe what the x-axis labels refer to (see changes on pp.XX).

**Reviewer 1**: I am not a native English speaker so I might be wrong. However, I thought it was odd to use the term "unaware psychological processes" in the Abstract. To the best of my understanding processes are not those with awareness. Minds have awareness. So minds can have awareness of processes. Similarly, I am not sure that the term "influence-aware trials" makes sense. But, perhaps it is the best abbreviated term to refer to "trials in which participants reported a priming effect."

**Authors**: We agree with Reviewer 1 that the phrasing “unaware psychological processes” was a bit strange, and have now revised this in the abstract. We opted to keep the term “influence-aware trials” because (i) we feel it is the most appropriate abbreviation, and (ii) the term “influence-awareness” has now been used elsewhere (albeit in a different context) since the submission of this manuscript (Sava, Payne et al., 2019).

**Reviewer 2**: This paper reports five experiments using retrospective self-report to measure whether participants are aware of being influenced by primes in the AMP. In each study, participants who exhibit greater priming were more likely to indicate that they were influenced by the prime. The authors then treat reported influence as a moderator, and find that the task appears to produce systematic and valid priming effects only among participants (or trials) where high levels of awareness are reported. They argue that this undermines the validity of the AMP as an implicit measure.

As the authors note in their literature review, this paper follows another paper by Bar-Anan and Nosek (2012) that took a similar approach to make similar claims. Those claims were rebutted by Payne et al (2013) and Gawronski and Ye (2014; 2015), who found that the evidence was consistent with a post-hoc confabulation account. That is, rather than accurately reporting the cause of their ratings, participants observed their responses and then reported whether they had been influenced (and if so, it must have been intentional). However, the authors argue that the present paper is different because whereas Bar-Anan and Nosek had participants complete an AMP and then give a holistic retrospective rating of whether they were influenced, the present paper asks participants to respond to the AMP on each trail, and then judge whether they were influenced by the primes on that trial. They argue (but do not provide any evidence) that the trial-by trial method is not vulnerable to post-hoc inferences.

**Authors**:

We would contend that our method is much more similar to Payne et al. (2013) than to Bar-Anan & Nosek.

We entered into an email correspondence with Prof. Payne after the initial round of review. He kindly provided us with feedback on Experiments 6 and 7. We are aware, based on this correspondence, that Prof. Payne believes that the influence-awareness measurement in Experiment 6 is still susceptible to post-hoc confabulation, and that Experiment 7 cannot speak to whether the mechanism driving AMP effects is one of misattribution.

Nevertheless, we feel that both Experiment 6 and Experiment 7 provide further evidence that AMP effects are based on participants’ awareness of the influence of the primes on their evaluative responses. We would also contend that these experiments demonstrate that this method of assessment is not susceptible to post-hoc inferences, as the influence-awareness response is emitted before participants even have an opportunity to evaluate the target overtly (Experiment 6) or covertly (Experiment 7).

**Reviewer 2**: However, a fundamental problem for this paper is that this method is still a retrospective self-report. Trial-by-trial retrospective reports are used routinely to demonstrate post-hoc inferences of the type in question here. For example, Aarts, Custers, & Wegner (2005) used a trial-by-trial retrospective judgment to show that participants often falsely claim authorship over "decisions" made by a computer. Many other studies have used a similar immediate retrospective judgment (e.g., Wegner's I Spy study, Wegner & Wheatley, 1999).

**Authors**: See above.

**Reviewer 2**: Another paper using immediate trial-by-trial retrospective reports to demonstrate post-hoc confabulations is Kühn and Brass (2009) which, strangely, is cited in this paper as evidence that unambiguous and immediate retrospective reports are likely to be accurate. In fact, that paper found that when people made impulsive errors in a stop signal task they often falsely claimed to have intentionally decided to make that choice. Kühn and Brass conclude, "Our data support the retrospective account of intentional action," (p. 12) based on the same kind of immediate retrospective reports used in this manuscript.

The similarity between the immediate retrospective reports used in the present studies and the holistic retrospective reports used in Bar-Anan and Nosek (2012) should be clear from the fact that they are correlated so highly (r = .78).

**Authors**: We apologize for this error on our behalf. This was a case of a misplaced citation on our part. The intended citation was in fact “Retrospective and Concurrent Self-Reports: The Rationale for Real-Time Data Capture” (Schwarz, 2012). We have now revised the manuscript to include the correct citation.

**Reviewer 2**: So why is it such a problem that the studies used retrospective self-reports that are vulnerable to post-hoc inferences? Statistically, this is an error known as "post-treatment bias" (Coppock, 2019; Montgomery, Nyhan, & Torres, 2018). It occurs when researchers use a variable that is affected by an experimental manipulation as a covariate or moderator to make inferences about the experimental effect. This creates a confound between the post-treatment variable and the experimental effect on any other outcome. In other words, this is a form of non-independent selection of the same form criticized as "voodoo" correlations by Vul et al., (2009). Concretely, if larger priming effects (the experimental effect of primes on ratings of pictographs) lead subjects to claim they are aware of the influence, then reported awareness can't be used as a meaningful moderator of the priming effect.

**Authors**: We hope that the reviewer feels that at the very least Experiment 7 can effectively rule out such a “voodoo correlation”.

**Reviewer 2**: Another way to look at this problem is that all of the analyses depend on the correlation between reports of awareness and the priming effect. The authors interpret their findings as evidence that people who show systematic priming effects have disregarded the instructions and intentionally rated the targets consistent with the primes. That is, aware and intentional ratings cause the priming effects. But all of the findings are just what the misattribution account predicts also. The misattribution account says that it is difficult to disentangle affective response to the primes and targets, so subjects often mistake the source of the affect as the pictograph target when it is actually the prime. (A misattribution by definition can't be made with awareness or intention). Participants can observe their own behavior and notice if they are responding in prime-consistent ways. If so, they can report afterward that they were influenced by the prime (see Payne et al, 2013 for the same argument). This means that when priming effects are larger, subjects should report more influence of primes. If you divide subjects into those that reported large influences and those who didn't, then those who did not report influence won't have much priming because they have been selected to be that way. So these studies do not distinguish between the misattribution account and the authors' intentional/aware account at all.

**Authors**: We would like to respond to Reviewer 2’s various points in turn.

Point 1. XXX

First, just to clarify, we are not saying that participants “disregarded the instructions and intentionally rated targets consistent with the primes”. What we do say is that participants were aware of the influence of the targets on their evaluations of the primes, and that this influence may have occurred either intentionally or unintentionally.

Point 2. XXX

Second, given that a “misattribution by definition cannot be made with awareness” (as the reviewer states), and that our seventh Experiment in particular eliminates the possibility that participants are confabulating reasons for their evaluations post-hoc, surely this experiment (if not any of the other) represents evidence that the AMP is not driven by misattribution? At the very least, we believe our findings greatly contribute to this conversation. Again, we hope that the reviewer considers our new studies to demonstrate that our results are not easily explained by post-hoc confabulation.

Point 3. XXX

Third, establishing the nature of mental mechanism which is involved in producing the AMP effect (i.e., misattribution or some other mechanism) is of secondary importance to our results. As we outlined to the Editor and Reviewer 1, and make clear in our revised manuscript, our primary aims in this paper were to (a) determine if the AMP effect is implicit in the sense that previous research claims it is, and (b\_ determine if the AMP effect represents a valid measure of evaluations.

**Reviewer 2**: A related problem is that the authors confuse correlation for causation throughout the manuscript. When using reported awareness as a predictor or moderator of the priming effects, they routinely use causal language to say that awareness "drives" the priming effect. In fact, they say the priming effect was "driven by" aware subjects 142 times in the manuscript. If each time, the authors instead correctly wrote that larger priming effects were correlated with subsequent reports of awareness, the problems would be more transparent.

**Authors**: See our reply to the Editor’s XX comment and Reviewer 1’s XX comment. Briefly,

**Reviewer 2**: Experiment 2 found that reports of awareness were correlated with priming effects on a previously completed separate AMP, and Experiment 3 found the same thing when the other AMP measured attitudes on a different topic. The authors say that this pattern can't be explained by post-hoc confabulations, but it clearly can. These effects also follow from the misattribution account. All implicit tests are indirect tests: they measure evaluations by how the evaluation perturbs performance on some primary task. This means that scores on implicit tests are influenced not only by the evaluation of the attitude object but also by performance on the primary task. This has been known for many years and is why much has been written about how implicit tests are not "process pure" (Jacoby, 1991; Payne, 2001). Various modeling approaches, such as multinomial models (e.g., process dissociation, quad model) have been developed to deal with this, including a multinomial model of the AMP that estimates component of performance by separating evaluations of primes from the likelihood of making misattributions (Payne et al., 2010). These findings simply show that individuals who make more misattributions show larger priming effects across different AMPs and that they also report being influenced by the primes. Again, it's just a correlation with a retrospective self-report. And it is predicted by the misattribution account of the AMP.

**Authors**:

Given that we now provide evidence to suggest that our results are not easily explained by inferences based on retrospective report (Experiments 6-7), and also evidence that misattribution may not underlie effects seen here, we contend with the comment that the consistency of AMP effects across domains within participants can be simply explained by a propensity to misattribute.

It seems to us that Reviewer 2’s explanation takes misattribution as a premise for the AMP effect rather than one possible conclusion. And that we are interpreting the same evidence as supporting two very different perspectives. We repeatedly asked Reviewer 2 to propose a study that would allow us to falsify a misattribution account and never received an answer.

In our paper we simply entertain a different premise for the same conclusion: what would an explanation of this effect look like if misattribution wasn’t the mechanism behind these effects? An explicit account would say that participants are simply prone to showing AMP effects or not in general, and that this is a function of the participants’ tendencies to intentionally and with awareness use the primes to inform how they respond. So both the explicit and misattribution accounts provide plausible explanations.

However, we now have evidence that AMP effects are greatly moderated by influence-awareness, and at least some of this evidence (i.e., Experiment 7) is not easily-explained by post-hoc confabulation. By the reviewer’s own position, misattribution cannot with awareness. Thus, our evidence now suggests that the explicit account is likely a better candidate to explain these findings.

**Reviewer 2**: In the introduction the authors attempt to argue against some of the previous points made in the exchange between Bar-Anan and Nosek and Payne et al (2013) and Gawronski and Ye (2014, 2015). First, they argue that it is problematic that the AMP defines what is intentional and unintentional by the instructions, and they note that sometimes subjects don't follow instructions and instead incorporate information that the researchers instruct them to ignore (p. 11). Subjects sometimes do this, of course, but the question at issue is why. Unintentional effects of primes on judgments is one reason they do so, although there are of course other reasons. Nonetheless, using instructions to define intentional responding is not a weakness. In fact, virtually every task that aims to measure performance by accuracy and errors must use instructions to define task goals and therefore what is accurate or error, and what is intended vs. unintended responding. For example in the Stroop task, experimenters must use instructions to tell subjects to name the font rather than read the words. Responses that diverge from the task goal (which is set by instructions) define automatic or unintentional behavior.

**Authors**: The reviewer states that “virtually every task that aims to measure performance by accuracy and errors must use instructions to define tasks goals and therefore…what is intended vs. unintended responding”. Is this true? If so, this can create a paradoxical situation where intentional behaviour is considered unintentional. If I instruct participants “rate the target, ignore the prime”, and participants intentionally rate the prime anyway because they ignore the instructions, then by the reviewer’s logic this should be considered unintentional behaviour (since the participant has diverged from the task goal). I can also, by this logic, construct a procedure where all behaviour is unequivocally unintentional by specifying an impossible response is required. If I provide a Likert scale from 1-7 and specify in my instructions to the participant that they must respond by clicking the number “8”, then any response they elicit must by this definition be unintentional, since their response deviated from the stated instructions.

Automaticity when defined in terms of goal-relevance is always defined based on participant goals (e.g., see Moors & De Houwer, 2006). Ultimately, we believe the reviewer here is erroneously assuming that the experimenter-specified goal is identical to the participant-derived goal. If the participant reads the instruction of “ignore the prime, evaluate the target”, but takes from this (via also seeing the procedure, appreciating that they are in an experimental context, etc.) that what the experimenter *really* wants is for them to evaluate the primes, then the goal of the participant becomes to respond based on what they believe the experimenter *really* wants. Of course, this may not be the case: some participants may have the goal to respond in opposition to what the experimenter wants. Some participants may have the goal to finish the experiment as quickly as possible. Some participants may have the goal to earnestly respond in accordance with the experimental instructions. However, we cannot necessarily infer that the goals of the participant are the same as those intended for them to have by the experimenter.

Reviewer 2: Moreover, the paper never offers an explanation for why large subsets of subjects would choose to ignore the task instructions and instead intentionally rate the primes.

**Authors**: We are not arguing nor have we ever argued that participants intentionally rate the primes in our previous or current manuscript. As such this argument does not apply to the current manuscript.

However, in the spirit of good faith and openness, we can *speculate* that there could be several reasons why participants might do so. For instance, (i) some participants could “read between the lines” and respond based on the perceived demands of the experimenter (i.e., conversational norms); (ii) some participants could simply fail to thoroughly read the instructions, and then midway through the task “figure out” what they should do (and the most obviously afforded task-goal is to evaluate the primes); (iii) participants initially try to adhere to the task’s instructions, but then quickly find it difficult to rate the Chinese characters and easier to intentionally use the primes as a source of inspiration. In this way some participants persevere with rating the primes whereas many others simply switch to the easier goal of rating the primes instead of the targets; (iv) participants aren’t particularly invested in the experiment and just do whatever is easiest and most entertaining for them from the outset, which either involves giving ratings of the targets or responding randomly. All of the above explanations can both (i) account for why participants ignore the instructions, and (ii) account for why this happens only amongst a subset of participants. Note that all of these explanations seem plausible but are nevertheless post-hoc justifications for our effects, so we reluctant to afford them any strong theoretical value.

**Reviewer 2**: Next, they argue that there are "statistical issues" in the Payne et al. (2013) paper. This section is full of factual errors. The paper says, "the authors found that the difference scores on 'unintentional' AMP and explicit race measures was larger than the difference between scores on the 'intentional' AMP and explicit race measures, and used this dissociation as evidence of unintentionality in the traditional AMP." But the Payne et al (2013) paper did no such thing. There were no comparisons between the size of difference scores with explicit measures.

**Authors**: We thank Reviewer 2 for highlighting these issues. On the one hand, we acknowledge that our characterization of the Payne et al. study was indeed factual incorrect in a number of instances. We sincerely apologize for those errors and have revised the manuscript to correct this (and other) such issues (see our reply to Reviewer 1’s XXX comment and revision in the manuscript on pp.XX).

On the other hand, we argue that there are statistical issues present in Payne et al. study. STATE THESE CLEARLY HERE.

Our criticisms of these statistical issues is now prefaced with accurate descriptions of the specific analyses conducted, as well as the related inferences drawn from these analyses (see changes on pp.XX).

**Reviewer 2**: Next the manuscript says "Critically, however, the inference that 'intentional' AMP effects were "more affected" (p. 381) by the race of the prime than 'unintentional' AMP effects was never directly addressed in any of their other analyses…" and then go one to say we should have tested an interaction rather than reporting that an effect on one version of the test was significant and the other was not. But the present authors are entirely mistaken about the analyses we reported, and so their criticism is uninterpretable. That study examined the associations between two forms of the AMP (an indirect version in which subjects judged the pictograph targets and a direct one in which they were instructed to rate the primes) and impression judgments of a black or white target character (we examined main effects and interactions in a regression framework). And we tested the effect of seeing the black target character versus the white target character on indirect and direct AMP tasks. The hypothesis tested was that when people intentionally rate the primes their responses will be more reactive than the indirect version to the task they just completed. It is not clear how to respond to the statistical issues raised in this section given that the errors make it difficult to know what the authors are talking about.

**Authors**: We hope the reviewer will be better able to respond to the statistical issues now that we have clarified our meaning.

**Reviewer 2**: Finally, the authors note as a "conceptual issue" that in the 2013 study, "divergence from explicitly endorsed attitudes does not necessarily mean that the AMP captures unintentional behavior. Measures that are structurally dissimilar can show apparently unrelated effects due to the differences inherent in the measure" (p. 14-15). In the 2013 study, direct and indirect forms of the AMP were used, in which everything was held constant except the instruction to rate targets versus to rate primes. These direct vs. indirect forms of the task are actually the most structurally matched implicit-explicit comparison in the literature on implicit attitudes (we proposed this method in a 2008 paper entitled, "Why do implicit and explicit attitudes diverge? The role of structural fit"). So I don't know what the authors are talking about here.

**Authors**: We may have been unclear in the original manuscript: this criticism was not levelled at the 2013 paper, but rather at the broader literature in general which has used such divergences between structurally-dissimilar measures as evidence for divergence between constructs. Indeed, we believe that the 2013 study is the only study to date which has avoided this specific issue in the context of the AMP’s automaticity. We did not, by any account, intend to imply that the 2013 suffered from this issue.

We have revised the manuscript to clarify this point (see changes on pp.XX).

**Reviewer 2**: I don't normally comment on silly titles, but the reference to The Emperor's New Clothes implies not just that previous research with the AMP is mistaken, but that researchers in the field are fools for believing something that is obviously nonsense. This implication is gratuitously insulting, and suggests a lack of insight into the strength of one's own evidence.

**Authors**:

**Reviewer 2**: For the reasons described above, I don't believe the data reported here distinguish between the misattribution account and an aware/intentional account of AMP effects. I also don't believe they provide any new insight beyond the previous Bar-Anan / Payne / Gawronski exchange. Due to the basic error in using a retrospective self-report to make inferences about the causes of the priming effect that preceded it, I do not believe the data warrant publication. In retrospect, however, I am aware that it is possible that I may be biased.

**Authors**:

Signed,

Keith Payne

References

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. Consciousness and cognition, 14(3), 439-458.

Coppock, A. (2019). Avoiding Post-Treatment Bias in Audit Experiments. Journal of Experimental Political Science, 6(1), 1-4.

Kühn, S., & Brass, M. (2009). Retrospective construction of the judgement of free choice. Consciousness and Cognition, 18(1), 12-21.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. American Journal of Political Science, 62(3), 760-775.

Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A process model of affect misattribution. Personality and Social Psychology Bulletin, 36(10), 1397-1408.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on psychological science, 4(3), 274-290.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. American psychologist, 54(7), 480.