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

**Editor’s comments**

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 want to thank the Editor and Reviewers 1-2 for their constructive feedback. We took that feedback seriously and used it to substantially revise our manuscript. It also led us to carry out three new studies that directly speak to the issues raised during the initial round of reviews. We believe these empirical additions, as well as the revisions to our manuscript, address the original concerns and have resulted in an even stronger paper.

**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. Many of these issues are no longer relevant as the content that was flagged is no longer present in this substantially revised version of the manuscript. That said, we still believe that one study on the unawareness of the AMP effect does contain a number of conceptual, statistical, and methodological concerns. We have done our best to accurate summarize and represent those concerns in a fair and balanced manner (for more see the revised introduction).

**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 suggestions we have amended any reference to the idea that prime influence awareness “drives” AMP effects in the revised manuscript. We now instead state that the AMP effect is “strongly moderated by” awareness of influence of the primes, which removes the causal element of our language while still conveying a conditional relationship between AMP effects and awareness.

We also carried out two new experiments that sought to directly test the idea that “all the observed effects in this paper are consistent with the implicit misattribution plus some post-hoc justification”. In our original submission, we only assessed awareness in a *retrospective* manner (i.e., after every AMP trial where the target was evaluated, and also after the task was complete). Although such a measure captures awareness mere milliseconds after a target is evaluated, it is nevertheless possible that performance is still a post-hoc confabulation on the participant’s behalf.

In two new experiments (7 and 8) we sought to control for this concern. Specifically, we created a *prospective* measure of awareness, where we asked participants to indicate if the prime will influence their response to the target and asked this either (a) before the target was evaluated (Experiment 7) or (b) before the target was even presented (Experiment 8). In this way, performance on the influence awareness measure could not be a post-hoc confabulation because it occurred *before* the target was evaluated or even presented onscreen. In both of these studies we obtained the same pattern of findings as we did in our previous six experiments. We explain in the General Discussion why our findings are highly incompatible with “implicit misattribution plus some post-hoc justification” perspective, and discuss what it would take to retain the idea of misattribution if one opted to do so (see Experiments 7-8 and the General Discussion).

**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**: We agree with the Editor that the ‘implicitness of AMP effects’ is a topic that has already been discussed elsewhere in the literature. In the revised manuscript we now make it clear that much of that work has focused on one aspect of implicitness (unintentional), whereas we are focusing on another less studied aspect (unawareness). We have revised the introduction in this version of the paper to better clarify our position and how we set about addressing those issues (see revised Introduction). We have also altered the title.

**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 have revised our paper to speak to the Editor’s points in the following ways:

1. Our revised introduction unpacks the issues with past work on the implicitness of the AMP effect, avoids misunderstandings or misrepresentations, and clarifies the ways in which we addressed and improved upon what came before us.
2. We included a new, high-powered, replication attempt of a key study in support of the unawareness of AMP effects (Payne et al. 2013; Experiment 3). Results indicated that the original study (and thus claims) failed to replicate, further reinforcing the need for the role of awareness of AMP effects to be re-examined.
3. Two new empirical studies were carried out that extended our original analyses from retrospective to prospective awareness, and thereby providing a strong rebuttal of the post-hoc confabulation idea forwarded by Reviewer 2.
4. Analyses illustrating that a small subset of participants are responsible for the majority of variance in group-level AMP effects, and that these participants tend to be the same subset of people who are responsible for effects across different AMPs and across attitude domains. This represents another critical finding relating to the AMP: that AMP effects are not reflective of the population *in general* but rather a specific subset of people (the highly influence aware). To our knowledge this is the first finding of its kind, and represents an important contribution to the literature in terms of moderators of the AMP effect (and perhaps implicit measure effects more generally; for a detailed discussion see the General Discussion).

**Reviewer 1’s comments**

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**: We took Reviewer 1’s comments seriously and decided to carry out two new empirical studies (see Experiments 7-8). This time we assessed awareness in a *prospective* rather than the *retrospective* manner used in Experiments 1-6. Participants were asked to indicate if the prime stimulus *will* influence their response to the target, and asked this either (a) before they had to evaluate the target, or (b) before the target stimulus was even presented onscreen. Once again we obtained the exact same pattern of findings as we did in those prior studies which used a retrospective measure. Given that awareness was assessed before either the evaluation was emitted or the target was presented, we cannot see how those effects represent a post-hoc confabulation (given that there is nothing to confabulate in such a design) or be driven by misattribution (seeing as the participant is declaring that the prime is *going to* influence their response to the target before they even see that target). Such a statement runs contrary to the very notion of misattribution as it has traditionally been defined.

Finally, Reviewer 1 asked whether awareness might not occur at all unless directly prompted (as was the case in the IA-AMP). We directly speak to this point in Experiments 2-8. That is, 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 this is what occurred in all seven of our studies which directly tested this relationship (including one study where the same ‘skip-AMP’ paradigm was used as in Payne et al., 2013). Moreover, Experiment 5 also demonstrated that the relationship between influence awareness and AMP scores is *bidirectional*, insofar as AMP effect sizes predict how influence aware one will later report being, and how influence aware one is at an earlier point in time will predict the later magnitude of their AMP effects.

Thus we think it unlikely that awareness was simply a post-hoc irrelevance. If anything it appears to be a core factor predicting the size and predictive validity of 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, it may be that valence is misattributed from the prime to the target, and that even though people know this is happening they still do so, either 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 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, we want to be clear. All of these possibilities are post-hoc justifications for the findings we report in our paper and not *a priori* claims made in previous empirical or theoretical papers. As far as we are aware, past work on the AMP effect and misattribution as a mental process do not claim that misattribution occurs *prospectively* and with *awareness* (see Experiments 7-8). 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 acknowledged that an existing theoretical concept does not – as it stands – predict nor explain the pattern of findings we obtained across eight pre-registered, high-powered studies with different attitude domain (political vs. generic), variants of the AMP (standard, Mann et al., IA-AMP), and awareness measures (retrospective and prospective)

Rather what is being offered here is a post-hoc adjustment to a concept in order to modify that concept so that it can be retained in some form. But any such adjustment (i.e., that misattribution occurs prospectively with awareness) 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**: As noted above, Experiments 7-8 were designed to address this alternative interpretation of our original findings, and our results repeatedly contradict such an account. That said, we acknowledge this possibility in the discussions of those studies to provide a reminder to readers that alternative (post-hoc) perspectives are 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**: As we note in our response to the Editor, we have toned down the causal language substantially in the manuscript and have removed mention of the word ‘drive’ throughout. We now state that AMP effects occur under the condition of awareness/that AMP effects are moderated by influence-awareness.

**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 have revised the paper in order to make it clear what it is that we actually are claiming, and what it is that we are not claiming. We *are* claiming that the AMP effect is not implicit in one way that many have claimed it is (i.e., unawareness). We believe that our findings systematically and clearly speak to this issue.

We are not claiming that the AMP effect does not meet any automaticity condition. For instance, target evaluations are typically emitted quickly in the task (thus satisfying the speed condition). To our knowledge the ‘goal-directed’ nature of the effect has not been investigated, so it may be the effect is implicit is that sense too. We simply argue that ‘implicit’ is an umbrella term of a set of automaticity conditions, and that the AMP effect meets some of these criteria (speed) but not others (awareness). In this way it is ‘implicit’ in some ways and not others.

**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. We have revised the introduction to include mention of this aspect of Payne et al.’s (2013) Experiment 2 design.

**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**: See our previous response.

**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 respectfully disagree with the Reviewer and maintain that our original argument here holds. The original version of the skip-AMP provides incomplete data insofar as it requires participants to *either* evaluate the target (i.e., provide evaluative information) *or* indicate that their response would have been influenced by the prime (i.e., influence aware information). But never both. As such, it is impossible to directly compare performance on trials where people indicated that they were influenced to those trials where they reported no such influence. Without both pieces of information, it is difficult to determine what impact influence-aware trials have on the AMP effect, and if this impact is comparable to, or greater than, that of the non-influenced trials.

As Reviewer 1 suggests, one can make this comparison *between* participants, by comparing the standard AMP effect to the skip-AMP effect. But we believe that a stronger demonstration is one that is made within participants where both pieces of information are obtained from the same person, rather than across different individuals. It is also worth noting that this comparison failed to emerge in the original study and failed to replicate in our direct replication attempt (Experiment 1). Thus we are reticent to place too much strength in it.

Reflecting back on the Reviewer’s comment, we realize that we differ in how informative a non-significant difference is between the skip-AMP and standard AMP. We agree with the Reviewer that *finding a statistically significant difference* between the skip AMP and standard AMP would represent informative evidence (namely, it would provide evidence for the role of awareness in the AMP). However, we contend that the *absence of a statistically significant difference* between those AMPs does not warrant the inference which Payne and colleagues made (namely, that “this opportunity for selective responding did not…reduce the priming effects”).

**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).

**Authors**: We have now simplified our point in order to better clarify our argument: that Payne et al. inferred statistical equivalence on the basis of the absence of statistical differences. Specifically, the point we made now simply states: “Yet this conclusion is also questionable given that non-significant statistical differences between two means does not necessarily imply that they are statistically equivalent (Lakens, Scheel, & Isager, 2018; Quertemont, 2011). As such the original inference drawn was not supported by the analyses conducted”.

**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**: Reviewer 1 asked “why would there be more priming when the prime was neutral rather than of clear valence”. However, it was not the case that there was more priming but rather more *skipping* on such trials. We now acknowledge Payne et al.’s original claim about these skipping responses and offer an alternative explanation according to the explicit account (see p.8-9).

**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**: As formerly noted, we have now substantially revised the introduction, and made it clear that much of the previous work on the implicitness of AMP effects centers on the issue of intentionality, and that relatively less work (with the exception of Payne et al., 2013, Experiment 3) focused on the issue of awareness. We hope the Reviewer now finds that we characterize this experiment more fairly and accurately. Additionally, we hope that our simplification of our two issues with this experiment (inferring statistical equivalence from an absence of statistical differences, and the inability to examine influence-aware vs. non-influence aware responses) helps to more clearly express our issues with the original experiment.

**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**: Given our substantially revised introduction, and the fact that Gawronski and Ye’s study dealt with the intentionality of AMP effects (and not the awareness of AMP effects) we have now omitted this study from our introduction.

**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. We also agree with the reviewer that at present it is unclear how the AMP measures up to other measures within psychology in terms of its heterogeneous validity across participants. We agree that this issue is worth exploring in the context of other (implicit) measures and explicitly address this in our revised General Discussion. Unfortunately, there is currently little-to-no data available now that would allow us to make such a comparison.

**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**: We realized that the structural validity issue is a separate (and substantive issue), and one that requires far more time and space to unpack than we have in an already long paper. We have therefore removed this section from the current paper and are now writing it up as a separate short-report for publication elsewhere. We thank Reviewer 1 for his comments and have incorporated them into that short-report.

**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**: We recognize that our phrasing of this point (that the primes *only* exert influence on ratings when participants are highly-influence aware) was too strong a position. We have therefore rephrased the manuscript to reflect the fact that the primes appear to *predominantly* exert influence on ratings when participants are highly influence aware. We now state that although the non-influence aware trials do contribute to the magnitude of AMP effects, their contribution pales in comparison to that of the influence aware trials.

**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**: In line with Reviewer 1’s suggestion we now provide more information about the AMP and its procedural parameters (see p.4-5).

**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**: We have revised the manuscript in order to better to clarify this point (see changes on p.19).

**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**: We have included the descriptive statistics and plots as requested (see Supplementary Materials).

**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. 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 meta-analysis section).

**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 elaborated our discussion of the magnitude of moderation in the meta-analysis section. For example, on page 46 we now state:

“Results demonstrated that a large proportion of the variance in AMP effects was attributable to the influence awareness rate between participants, *B* = 0.52, 95% CI [0.48, 0.55], β = 0.60, 95% CI [0.56, 0.64], *p* < .001. Recall that the AMP effect is the difference in evaluations on trials involving positive versus negative primes, and can range from 0 (evaluations unrelated to prime valence) and 1 (all evaluations congruent with primes). The model intercept was *B* = 0.14, 95% CI [0.12, 0.16], β= -0.01, 95% CI [-0.07, 0.05], *p* < .001. At the two extremes, in participants who report being aware of the influence of the prime on their evaluations on 0% of trials, the estimated marginal mean AMP effect on the IA-AMP was therefore 0.14. In contrast, in participants who report being aware of the influence of the prime on their evaluations on 100% of trials, the estimated marginal mean AMP effect on the IA-AMP was 0.66. The AMP effect was therefore estimated to be three times larger in fully influence aware participants than fully non-influence aware participants.”

With regard to moderation of the IA-AMP effect at the trial level, this is most clearly illustrated by the estimated marginal means for each level. These are presented in Figure 4 (page 46).

**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**: This section was removed from the paper during revision and no longer applies.

**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 no longer make use of the term ‘drive’ in the paper, and instead we make the specific claim of statistical moderation. We have also revised this section of the paper to clarify precisely what it is that we are claiming (see p.27-28):

“A significant effect emerged in both the standard AMP (OR = 3.10, 95% CI [2.87, 3.35], p < .001) and IA-AMP (OR = 4.66, 95% CI [4.30, 5.05], p < .001). At the trial-by-trial level, IA-AMP effects were moderated by influence aware trials, OR = 20.65, 95% CI [17.10, 24.94], p <.001. At the group level, IA-AMP effects were predicted by the influence awareness rates of participants, B = 0.44, 95% CI [0.34, 0.54], β = 0.56, 95% CI [0.44, 0.68], p < .001.”

**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 results for the replication analyses in each experiment. We hope the inclusion of these statistics, coupled with the statement that the effects replicated, and the meta-analytic effects in the meta-analysis section, will be sufficient.

**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**: We thank the reviewer for this useful insight – we had not considered that this does indeed provide additional evidence that the effect is robust for presentation order. The order of the tasks was fixed in order to maximize congruence with the previous experiment: participants always completed the politics IA-AMP first followed by the valence IA-AMP. We have added the following to the manuscript on page 38;

“It is also useful consider the implications of these results in terms of temporal order rather than domain. Although it was not part of our original research plan, these results also suggest that the temporal order of the tasks, and therefore the order of assessment of the AMP effect versus the influence rate, does not matter. Participants always completed the politics IA-AMP first and the valence IA-AMP second. The influence rate in the politics IA-AMP (completed first) predicted the absolute magnitude of the valence IA-AMP (completed second), B = 0.46, 95% CI [0.36, 0.55]. Equally, the influence rate of the valence IA-AMP (completed second) predicted (or more accurately 'postdicted') the absolute magnitude of the politics IA-AMP (completed first), B = 0.49, 95% CI [0.38, 0.50]. The very similar estimates and strongly overlapping confidence intervals provide no evidence that order of presentation moderated the effect.”

**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 figures throughout the manuscript to explicitly describe what it is the x-axis labels refer to.

**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’s comments**

**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 trial, 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**: Based on the comments of Reviewers 1-2 we decided to conduct three new studies. First, we carried out a high-powered, pre-registered replication attempt of Payne et al. (2013; Experiment 3). Results indicated that the original findings did not replicate insofar as standard AMP effects were larger than those obtained from the skip-AMP.

Second, and more importantly, we carried out two new empirical studies (Experiments 7-8) that swapped the retrospective awareness measure for a *prospective* measure. Specifically, participants were asked to indicate if their response to the target stimulus will be influenced by the prime, and asked this before the target evaluation was emitted (Experiment 7) or the target stimulus was even presented (Experiment 8). In both cases, the same pattern of findings emerged as in our previous studies with retrospective measures (Experiments 1-6).

Post-hoc confabulation likely cannot take place in Experiment 7, and certainly not in Experiment 8 given that there is nothing to confabulate. In short, these new experiments provide evidence that the trial-by-trial method is not vulnerable to post-hoc inferences.

**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 our previous comment.

**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**: See our previous comment. Also 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 recognize that this criticism may be levied at Experiments 2-6. However, it does not apply to the newly added Experiments 7-8 that use a prospective measure. Additionally, the finding that the same pattern of results emerged in Experiments 7-8 as did on Experiments 1-6 further increase our confidence that the relationship between influence awareness and AMP effects is not a ‘voodoo correlation’ as Reviewer 2 claims, but rather a bidirectional relationship that holds across eight high-powered, pre-registered studies, with multiple versions of the AMP (standard, Mann et al., IA-AMP), attitude domains (politics vs. general attitudes), awareness measures (retrospective and prospective), and samples (general population vs. those with specific political orientations).

**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 once again point to the findings of Experiments 7-8 which are incompatible with a post-hoc or a misattribution account.

*Post-hoc confabulation*

As we mentioned above, a post-hoc confabulation perspective requires that a prime is presented, a target is evaluated, and only then is participants asked to report on their awareness (either after each trial [as in the skip-AMP and our Experiments 1-6] or at the end of the study). At this point in time their response on the influence awareness question is said to be a confabulation, insofar as they notice the correspondence between their response to the target and prime valence, and use this to justify their response on the awareness measure.

Yet we used a prospective measures in Experiment 7 such that a prime was shown, a target was shown, influence awareness probed, and only then was a target evaluation emitted. In this instance there was nothing to confabulate because the target response had not yet been emitted. Moreover, Experiment 8 avoids the possibility that confabulation is taking place due to a covert evaluative response, because the prime stimulus was shown, influence awareness measured, a target stimulus shown, and then a target evaluation emitted. In this design influence was assessed before the target stimulus was even presented onscreen, and so it would not be possible for participants to confabulate any kind of evaluation of the prime with evaluation of the target.

We recognize that Reviewer 2 could propose further adjustments to the post-hoc confabulation account in order to make it fit with our newest findings. But, as we mentioned in our responses to Reviewer 1, any such adjustments are themselves post-hoc, and should not be granted equal evidential weight as the pre-registered hypotheses and findings noted here.

*Misattribution*

Reviewer 2 notes that “misattribution by definition cannot be made with awareness”. However, in Experiments 7-8 participants *prospectively* said that they were aware of the prime and indicated that it *will* influence their target evaluation before they emitted that response (Experiment 7) or even saw the target stimulus (Experiment 8).

If one wants to explain AMP effects in these two studies in terms of misattribution, then they would need to allow for the idea that people are not only *aware* of misattribution but also able to *predict* that it is going to occur before a target is evaluated or a target stimulus is even presented onscreen. Yet such an approach runs contrary to how misattribution is traditionally defined (Schwarz & Clore, 1983), and would require a radical overhaul of the concept itself.

Thus we believe that the findings from Experiments 7-8 are inconsistent with the concept of misattribution (as traditionally defined until this point) and would require a significant (post-hoc) change to that concept in order to accommodate the outcomes reported here.

**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**: As we note in our reply to the Editor and Reviewer 1, we have removed mention of the term ‘drive’ and replaced it with the term ‘moderated’ throughout the paper.

**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**: We refer Reviewer 2 to our previous comments in the context of Experiments 7-8. Although we acknowledge the proposed criticism of Experiments 2-6, it does not apply to Experiments 7-8, nor does it explain why findings in those latter experiments are identical to those in the former.

**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**: This section has been removed during revisions to the manuscript and these claims are no longer made. We thank the reviewer for catching these issues.

**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 have not intended to argue at any point throughout the course of this work that participants in the AMP intentionally rate the primes. Rather our core point is that people are aware of the prime’s influence on their target evaluations, and that they are aware of this before they even encounter the target stimulus, before they rate the target stimulus, or after they rate the target stimulus. We have revised our manuscript throughout to avoid the implication of making such claims.

**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 recognize that our characterization of the Payne et al. study was incorrect on several fronts. 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 comment and the revised manuscript).

That said, we still content that there were a number of methodological, statistical, and conceptual issues in earlier studies which led us to re-examine the implicitness of AMP effects.

**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 apologize for any perceived mischaracterization in our original submission. We have substantially revised the revised manuscript as well as the description of this study (see revised introduction).

**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**: This comment no longer applies to the current manuscript as it was removed during the revision process.

**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**: This section of the manuscript has been removed during revisions and no longer applies. The title has also been altered.

**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:** We appreciate Reviewer 2’s comments. We now feel that the addition of our two new experiments addressing the question of post-hoc confabulation within our results, combined with our third new experiment attempting to replicate the findings of Experiment 3 of Payne et al., establish even more clearly the contribution which our work can make to the field and the question of awareness in the AMP.