Editor’s comments:

**Issue:** 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.

**Response:** We hope now, with the addition of Experiments 6 and 7 which seek to even further remove the potential influence of post-hoc confabulations in participants’ influence-awareness responses, that the Editor (and reviewers) will agree that our findings cannot be accounted for based on post-hoc confabulation.

We would also like to comment more generally by saying that, in our eyes, within the literature of awareness and/or intentionality within the AMP, there are almost invariably two or more competing explanations for effects which are observed, and there is essentially no instance in which both accounts cannot provide adequate explanation (i.e., even in those papers which show evidence for the implicit misattribution account and claim that the explicit account cannot accommodate for such results, the explicit account in fact *can* provide an explanation, as we also demonstrate below in our responses to both Reviewer 1 and Reviewer 2).

Because of this, we do not believe that it will be possible for one paper to give a definitive answer which supports one account and refutes the other, because both accounts can continually refine themselves in order to attempt to accommodate the new data (in our eyes, the implicit misattribution account has had to adapt to a much greater degree than the explicit account, however). However, what we now show here (when including Experiments 6 and 7) is that effects in the AMP appear to occur predominantly under the condition of awareness, which refutes the implicit misattribution account. We also show that this finding of awareness cannot be simply explained by post-hoc confabulation, since effects hold even before participants have had the opportunity to make an overt (Experiment 6) or even covert (Experiment 7) evaluation. The introductions of Experiment 6 and Experiment 7 now discuss (and address) the alternative explanations of our findings in detail. Additionally, we have tried where possible to provide (and respond to) alternative explanations or interpretations of our findings even beyond this. In addition to this, we have also now included comments from both Reviewer 1 and Reviewer 2 in the General Discussion (and our responses to them) in order to provide a more thorough discussion of alternative accounts.

**Issue:** 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.

**Response:** We appreciate both the reviewers and the Editor for catching us on these mischaracterisations. We did not intend to deliberately omit information: we instead focussed heavily on the issues that we perceive with this literature, and in the process failed to include details which were also relevant to the papers being discussed. We hope that our amendments now provide accurate characterisations of these papers. While we still believe the literature contains a wealth of statistical, methodological, and conceptual issues, we want to convey this argument while providing readers with accurate and fair information.   
  
**Issue:** 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.

**Response:** We have amended the reference to “driving” the AMP effect throughout the manuscript. We should note that our argument does not require that awareness “cause” the effects: rather, our argument simply requires that participants be aware of the influence of the primes on their evaluations of the target stimuli. We believe that we demonstrate this across our five experiments, but we take the point which Reviewer 2 noted that these experiments do not rule out the possibility of post-hoc confabulation. However, our 6th and 7th Experiments now directly 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 contribute greatly to the debate around the automaticity of AMP effects.   
  
**Issue:** 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.**

**Response:** While we agree that these issues have been hashed out in previous back-and-forths, the methodological, statistical, and conceptual issues which are prevalent in this literature make it difficult to come to a conclusion on the weight of evidence for either account. This was the origin of this manuscript in the first place: our observation of an ever-increasing number of conversations amongst researchers (including ourselves) who found the extant literature insufficient to know whether or not the AMP can qualify as “implicit” in the sense of unaware. With the addition of the two new experiments, we hope the Editor now finds that this manuscript can do as we originally intended: to provide a methodologically-sound, large-N, multi-experiment investigation of the nature of awareness within AMP effects which provides a large degree of evidentiary weight by overcoming the shortcomings of previous work.

In terms of the title…

**Issue:** 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.

**Response:** 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, that we speak very 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.

Reviewer #1:   
**Issue:** 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.

**Response:** A number of criticisms from both Reviewer 1 and Reviewer 2 relate to this potential alternative interpretation (i.e., that awareness arises only after the participant has elicited a response). In order to rule out this alternative interpretation, we have conducted two additional experiments: one which assesses influence-awareness *before* the participant elicits their overt evaluative response, and one which assesses influence-awareness before the participant even *sees* the target stimulus. In both cases, our results demonstrate that the moderation of the AMP effect by influence awareness cannot be simply explained away by such a “voodoo correlation”.

**Issue:** 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.

**Response:** Indeed, we agree with the reviewer that a potential interpretation can be that misattribution is still the mechanism which underpins AMP effects, but that participants are aware of the misattribution and unable to correct it. We explicitly acknowledge this possibility in the “Do AMP effects reflect a misattribution process?” section in our discussion. However, in response to the reviewer’s point that “in the AMP, participants cannot avoid misattributing even if they suspect that it occurred”, we are explicit throughout the manuscript that the results of our experiments speak to awareness of prime influence, and *not* the intentionality of the effect. Additionally, as we discuss in the manuscript, the theoretical account of misattribution as it stands suggests that misattribution can occur only in the absence of both intention *and* awareness (and the absence of awareness is the more critical aspect of this). Our results show that an effect commonly purported to involve misattribution occurs with the awareness of the participants, which contradicts this theoretical account. The concept of misattribution may be retained in some modified form with the alternative explanation that we and the reviewer discuss, but this would violate misattribution as traditionally conceived and would require an overhaul of all previously-published work on misattribution to date.   
  
**Issue:** 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.

**Response:** 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.

**Issue:** 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.

**Response:** 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). However, we have amended our language around specifying that it is awareness itself that drives AMP effects in general (although we still believe this to be the case).   
**Issue:** 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).

**Response:** 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).

**Issue:** 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.  
**Response:** See our point above.

**Issue:** 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.

**Response:** Thanks to the reviewer for pointing this out. It is indeed true that initially we omitted a description of this finding in our initial discussion of the study. Indeed, upon reading the comments of both reviewers and the Editor, we realise that in the course of providing details of the flaws of studies, we failed to provide an overview of other effects observed in those same studies which were not flawed. This omission was not intentional on our part; we simply wanted to elucidate the methodological, conceptual, and statistical issues present in previous work, and failed to provide sufficient coverage of other effects in the process. We have amended descriptions of every study (where appropriate) in order to provide more comprehensive detail in relation to this work.

**Issue:** 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).

**Response:** 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 this analysis: “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.”. However, the authors cannot infer that direct ratings were *more affected* than indirect ratings on the basis of this, because the authors do not directly compare the two measures. In order to do this, 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 found an interaction effect between character condition and rating type, then this might allow 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, their inference here is invalid.   
  
**Issue:** 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. 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).

**Response:** We agree with the reviewer (and indeed, Payne et al.) that, had a reduction been found, this would have provided evidence for the requirement of awareness for AMP effects. Indeed, this was in part the basis for the rationale for our experiments, and we are therefore glad that the reviewer accepts this position (as, by extension, we suppose that he accepts our position that our evidence for reduced AMP effects in influence-unaware trials in our experiments demonstrates the salience of awareness in producing AMP effects).

However, the authors of the original paper made the inference that “Th[e] opportunity for selective responding did not eliminate *or* *even reduce* the priming effects”. The italicised claim is not substantiated by the analyses which were conducted by the authors (because the authors never directly compared the magnitude of the two effect sizes). Rather, the analyses which the authors reported merely demonstrate the presence of an AMP effect for both AMP versions. Thus, the claim “did not eliminate the priming effects” would be valid, but the position that effects did not reduce is not. If the authors wished to substantiate this claim, they would have needed 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 their inference is invalid.

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

**Issue:** 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.

**Response:** We accept the reviewer’s point that we erroneously omitted this major finding, and have now included it within the manuscript.

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 true, 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 added this discussion into the conceptual issues section of our introduction.   
  
**Issue:** 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.  
**Response:** We hope the reviewer feels now that our characterisations of this paper is more accurate. 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).

**Issue:** 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.

**Response:** 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).

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 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 never directly test this claim. Given recent work demonstrating that such bespoke manipulations are often poorly-validated (Chester & Lasko, 2019), the conclusions that can be drawn from this work are limited at best. The authors also committed the inferential fallacy of concluding a significant difference from a nonsignificant (“marginal”) test statistic. We have amended the manuscript to elaborate on these points.  
  
**Issue:** 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.  
**Response:** We are glad that the reviewer agrees with us that our evidence highlights that the AMP is not an equally valid measure across all individuals. Additionally, we would suggest that the AMP is commonly-used to make inferences about *automatic evaluation* in the general population. The reviewer suggests that it is better if a measure works well for a larger portion of the population: we agree. However, we would suggest that the AMP does not work as a measure of automatic evaluations *at all.* If AMP effects reflect the evaluations only a subset of individuals (which seems to be around 20% of participants, and our work represents the first contribution of quantifying this percentage), then it 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. Thus, inequality in the validity of the AMP across individuals clearly represents a massive barrier to scientific progress that wishes to study automatic evaluation. For a given participant, we are either studying evaluation or automaticity using the AMP – but likely never both. We have added this point within our discussion to reflect our position on this.

We are currently pursuing work on this a point you raised in another project – 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.   
  
**Issue:** 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).

**Response:**   
  
**Issue:** 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.  
**Response:**   
The claim the reviewer firstly alludes to (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 – we are instead 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). Additionally, we should flag that the analyses the reviewer highlights for the presence of AMP effects refer to analyses done at the trial-level of analysis, whereas the discussion here is focused at the participant-level of analysis (where effects are related but distinct).

Ultimately, our point is not that participants do not show an AMP effect. Our point is 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).   
  
**Issue:** 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.

**Response:** This has now been added.   
  
**Issue:** 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.  
**Response:** This has been amended, and indeed was initially unclear.   
  
**Issue:** 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.

**Response:** These have now been added.  
  
**Issue:** 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.

**Response:** The reviewer is indeed 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. We would like to flag for the reviewer also that outputting Word or PDF versions of these files is possible through the use of RMarkdown within our original analysis files.   
  
**Issue:** 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.

**Response:** This has been amended.  
  
**Issue:** 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).

**Response:** We have now added the mean evaluation of each prime type at each level of influence-awareness.   
  
**Issue:** 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.

**Response:** “Readily” indeed means intentionally (and the authors of the referencing paper explicitly investigated the intentional usage of to-be-ignored information. We have amended this sentence to specifically include the use of “intentionally” to avoid any potential ambiguities in what we are referring to.   
  
**Issue:** 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.

**Response:** We have now amended this.  
  
**Issue:** 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).

**Response:** We know include the descriptive statistics for this replication analysis in each experiment in the inclusion of multiple Tables. We do not specifically report the p values for the results of these analyses (they are available in the supplementary materials). However, we hope now that the inclusion of these descriptive statistics, coupled with the statement that the effects replicated, and the overall meta-analysis of the effects in the meta-analysis section, will be considered satisfactory to the reviewer.  
  
**Issue:** 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.

**Response:**  
  
**Issue:** 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.  
**Response:** We have amended the description of the figure to now explicitly describe what the x-axis labels refer to.

**Issue:** 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."

**Response:** We agree with the reviewer that the phrasing “unaware psychological processes” was a bit strange, and have now amended 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:   
**Issue:** 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.  
**Response:** We would contend that our method is much more similar to Payne et al. (2013) than to Bar-Anan & Nosek. We have had private correspondence with Prof. Payne, who very 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).   
  
**Issue:** 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).

**Response:** See above.

**Issue:** 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).  
**Response**: Frankly, this was simply 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 amended the manuscript to include the correct citation.   
  
**Issue:** 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.  
**Response:** We hope that the reviewer feels that at the very least Experiment 7 can effectively rule out such a “voodoo correlation”.  
  
**Issue:** 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.

**Response:** Firstly, we would like to clarify that we are not saying participants “disregarded the instructions and intentionally rated targets consistent with the primes”. It would be more accurate to characterise our position as saying 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.

Establishing the nature of mechanism which is involved in producing the AMP effect (i.e., misattribution or some other mechanism) is of secondary importance to our results (which are most relevant to (un)awareness and the validity of the measure as a measure of evaluations). However, 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, as well as conversations relating to influence-awareness and validity. Again, we hope that the reviewer considers our new studies to demonstrate that our results are not easily explained by post-hoc confabulation.   
  
**Issue:** 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.

**Response:** We have amended this throughout.   
**Issue:** 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.

**Response:** Given that we now provide evidence to suggest that our results are not easily explained by inferences based on retrospective report, 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.

We note that the reviewer’s explanation takes misattribution as a premise. 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.   
  
**Issue:** 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 wat 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.

**Response:** 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. In the words of Skinner, the rat is always right.

**Issue:** 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.  
**Response:** This is because we are not arguing that participants intentionally rate the primes, so this is not an argument particularly relevant to the current manuscript. Additionally, these discussions have been had elsewhere in the literature. However, we can speculate that there could be several potential explanations as to why, for example: (i) some participants “read between the lines” and respond based on the perceived demands of the experimenter; (ii) some participants generally 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 quite difficult to rate the characters – some participants persevere, while others switch to just rating the primes; (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.   
  
**Issue:** 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.

**Response:** Although we do believe that statistical issues are present within Payne et al. (2013), we recognise that our characterisation of the study was indeed factual incorrect in a number of instances. We believe we have now corrected our characterisation of this (and other) papers throughout the manuscript, and our statistical criticisms are now prefaced with accurate descriptions of the specific analyses conducted, as well as the related inferences drawn from these analyses. We gratefully thank the reviewer for picking up on these issues.

**Issue:** 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.

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

**Response:** 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 now amended this point to clarify precisely what we meant.

**Issue:** 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.  
**Response:**

**Issue:** 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.