Re: PSP-A-2021-1336  
Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness  
Journal of Personality and Social Psychology: Attitudes and Social Cognition  
   
Dear Dr. Hughes,  
   
I have received three expert reviews of the manuscript that you and your co-authors recently submitted to JPSP-ASC, titled “Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness” (PSP-A-2021-1336). I would like to take a moment to express my gratitude to the reviewers for their effort and attention in reviewing this manuscript, particularly during this difficult time. 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, the reviewers are somewhat split in their recommendations about the paper but, at their core, are in consensus about the strengths and limitations of the paper. All the reviewers sees at least a possibility of a novel contribution in this case, but at the same time each raises issues related to the clarity and accuracy of the inferences you make from the data and their implications for the AMP. My own reading of the work places me in agreement with this general assessment of your work by the reviewers, particularly with regard to the possibility that people can infer the influence of a prime from their own affective reaction. So, though I cannot accept this version of the paper for publication in JPSP-ASC, I invite you to revise and resubmit the paper after addressing all the concerns raised in the reviews.  
   
The reviewers clearly expressed their concerns and thus I will not reiterate them. The main meta-critiques I see are about the clarity and validity of the interpretations of your data - exactly and specifically what do these data say and not say about the psychology of the AMP? - and the adequacy of your discussion of the prior literature. On that point, it seems that there are a number of prior discussions in the literature around similar if not identical issues. Please be sure to clarify how these studies are different and can advance those discussions past their previous boundaries.

**Authors**: We sincerely thank the Editor for his comments and feedback. We have once again revised the paper in line with those comments as well as the feedback of the three reviewers. We have also paid specific attention to (a) our discussion of the prior literature, (b) how our work connects to, and extends beyond, that literature, as well as (c) the clarity and accuracy of our claims as well as their implications for the AMP (see our responses to Reviewers 1-3 as well as the revised manuscript).

**Editor**: Once the paper has been revised, submit it through the manuscript submission portal. Make sure to check the appropriate box in the portal to indicate that the paper is a revision rather than a first submission. If possible, I would like to receive your revision by 07/20/2021. If this is not feasible, please email our Peer Review Coordinator, Charlie Retzlaff, at the main editorial office (cretzlaff@apa.org) with an estimate of when you will resubmit. Longer timeframes are fine.  
   
Your resubmission must be accompanied by a detailed cover letter explaining which specific changes you made and which recommendations you did not follow and why. This letter should address all of the points raised in my decision letter plus any other major, non-redundant points mentioned by each reviewer.  
   
In closing, thank you for submitting to **JPSP-ASC**. I would also like to thank the reviewers for their service to the field. Their thoughtful comments and suggestions were very helpful in reaching my decision.  
   
I enjoyed reading this paper and I hope you decide to undertake the revision.  
   
Sincerely,  
Elliot T. Berkman  
Associate Editor  
Journal of Personality and Social Psychology: Attitudes and Social Cognition  
  
  
​

Reviewer #1: Signed: Yoav Bar-Anan  
  
1. In my previous review of an earlier version of this manuscript, my opinion was that the experiments report novel informative findings that should be shared with the rest of the scientific community. This strength has further improved in the present manuscript. The main finding has remained the same: people's immediate judgment whether priming occurred in a specific trial is related to the priming effect in that trial (Experiment 2-6). Further, the frequency of reporting the priming in one modified AMP is related to the size of the priming effect in a previous standard AMP, with similar (Experiment 3) or different (Experiment 4) primes, and in the priming effect of a previous or future AMP that had the same modification of judging the influence of the priming after each trial (Experiment 5). The results of Experiment 3 were also replicated (in Experiment 6) using a recent modification of the AMP that was supposed to improve participants' ability to follow the task instructions (and therefore, perhaps, reduce intentional rating of the primes). In the studies added to this version of the manuscript, the authors found that participants can predict whether priming would occur before judging the target (Experiment 7), and even before seeing the target (but after seeing the prime stimulus; Experiment 8). With that addition, it is now more difficult to argue that estimating the priming occurred in a particular trial is based on observing the congruency between the evaluation of the target in that particular trial and one's attitude toward the prime in that particular trial. The experiments in this manuscript are rich with great novel findings and I have no doubt that they would interest many readers.

**Authors**: We thank the reviewer for his kind words and assessment of our work.

**Reviewer 1**: Previously, I thought that the writing of the manuscript did not provide an optimal service to readers that would be curious to understand this research and its implications. This aspect has improved in this version of the manuscript, with the omission of many inaccurate or unclear sections and statements. However, for a strong and effective contribution, it would be important to improve this aspect of the manuscript further. In short, the writing is sometimes inaccurate, omitting important details, or unclear. Writing clarity is often quite subjective, and it is difficult to comment on all the places that clarifications and more precision could have helped. Therefore, I will first focus on main conceptual issues, and then provide a few examples for imprecisions that might prevent this manuscript from reaching its full potential.  
  
2. What do the results add to our knowledge about the AMP? The authors emphasized that their studies only pertain to awareness. However, there was already evidence that, if asked about it, some people show some retrospective awareness of the effect (Bar-Anan & Nosek, 2012, and replicated by Payne et al., 2013, Gawronski & Ye, 2015, and Weil et al., 2017). Therefore, I do not think that there is currently an open dispute about whether people can become aware of the priming effect, if prompted about it.

**Authors**: We agree with the reviewer that a number of studies have already explored the relationship between reports on retrospective (self-report) *intention* measures and AMP effect sizes (e.g., Bar-Anan & Nosek, 2012; Gawronski & Ye, 2014; Mann et al., 2019; Payne et al., 2013; Weil et al., 2017). We have included additional material in this version of the manuscript acknowledging such work (see p.6-8).

We also agree that a number of studies have explored the relationship between retrospective (self-report) *awareness* measures and AMP effect sizes (e.g., Bar-Anan & Nosek, 2012; Payne et al., 2013). We have included additional material in this version of the manuscript acknowledging such work (see p.6-8).

We have also revised the paper to make it clear that we are *not* arguing that there is a debate about whether people can report awareness of prime influence when prompted about it. Several studies empirically demonstrate that people are capable of doing this.

What we are arguing is that researchers differ in how they *interpret* those empirical findings (i.e., either as evidence that awareness plays a causal role in the AMP effect vs. is simply something correlated with those same effects).

**Reviewer 1**: The authors' findings go beyond previous studies by measuring awareness of the priming within each trial, by using that measure of awareness to predict priming in another AMP, and by measuring reported priming within each trial before the priming could occur.

**Authors**: We have added new material to the end of the introduction to better clarify how our studies extend on past work in this area (see p.11-12). Specifically, we now state:

“Taken together, our studies build on previous work in this area by investigating awareness in both retrospective and prospective ways. They do so using multiple measures (single post-hoc self-reports, trial-by-trial online measures), versions of the AMP (standard, Influence Aware version, Mann et al. version), and attitude domains (political, positive vs. negative). They explore the bidirectional (predictive) relationship between performance on one AMP and that of another (something that has not been examined in prior awareness work), and examine how awareness plays out at the trial, individual, and group levels. The result is a level of generalizability about awareness in the AMP that extends beyond that which has come before.”

**Reviewer 1**: These are all very interesting and important to know because they could inspire new accounts for the priming effect in the AMP, which would lead to further research about the present findings. For example, it would be interesting to understand what kind of trials are more likely than other trials to elicit reports of a priming effect, and what leads people to consistently show priming effects in different AMPs, and to be able to then report that priming effect.

**Authors**: We agree with the reviewer that these are interesting questions for future research. We have added new material to the General Discussion highlighting these and related questions (e.g., individual difference factors that contribute to why people respond as they do within and between AMPs) (see material on p.62). We have also included a new section on the theoretical perspectives stimulated by our findings in the General Discussion as well (see p.58-59).

**Reviewer 1**: Yet, unlike the authors, I am not sure that there are any current accounts for the priming effect in the AMP that are not compatible with the present findings. Payne et al. (2013) suggested that the awareness of the priming effect is only *retrospective*, inferred after the fact, rather than something that exists while the priming occurs. This still allows for misattribution to explain the effect, as well as other accounts that do not require awareness (e.g., occasional confusion). Contrary to the authors' arguments, I think that the novel findings do not rule out the possibility that in the standard AMP, when participants are not required to report whether priming has occurred, there is no awareness of the priming effect.

**Authors**: Reviewer 1 suggests that, in the standard AMP, when people are not required to report influence awareness they are not aware of the primes influence on their target evaluations. While certainly possible there are several aspects of our work that continue to make us think otherwise.

First, if people are not aware of the prime in the standard AMP, then why would scores on that measure consistently and strongly be *predicted* by influence awareness rates on a subsequently completed IA-AMP (see Experiments 3-4)? Performance on the former cannot have been perturbed by the latter given that the latter was completed always at a later point in time. Second, all other aspects of IA-AMP performances were identical to those seen in the standard AMP (e.g., similar effect sizes, etc.). These two points make us think that our results with the IA-AMP can be applied to the standard AMP.

In short, we agree with the reviewer that one cannot fully eliminate this possibility. But the repeated, convergent, and replicated findings across our studies strongly suggests that what is happening in the IA-AMP is also related to what is happening in the standard AMP too.

**Reviewer 1**: That is also true for most of the modified AMPs in the present research. In Experiments 2-6, participants might have inferred from each response that it was influenced by the prime stimulus. This is also true for Experiment 7, if we assume that participants knew what their evaluation of the target would be, when they would be allowed to communicate it at the end of the trial. In Experiment 8, participants might have inferred that priming is likely to occur based on the extremity of their reaction to the prime, perhaps because they knew that, in general, they don't have much of an opinion about the targets.

**Authors**: Again we agree with Reviewer 1. This is certainly a possibility. But it appears to be a post-hoc interpretation of our findings rather than an *a priori* account that makes such predictions before seeing the data and associated outcomes. For instance, we could find no mention of an inferential account of AMP effects in the literature when we looked.

To be very clear: we are not arguing *against* post-hoc interpretations of our findings. We agree that our work poses many interesting questions that could lead to new theories and ideas about AMP effects. What we don’t feel so comfortable with is assigning post-hoc explanations equal explanatory weight as *a priori* specified (and more importantly pre-registered) hypotheses.

That said, we have added new material in the General Discussion acknowledging the inferential account forwarded by Reviewer 1 as one possible post-hoc explanation for our findings (see p.58-59).

**Reviewer 1**: What about the correlation between the priming effect in the standard AMP in Time 1, and reporting the priming effect in a modified AMP in Time 2? It suggests that there are consistent individual differences in the tendency to show the priming effect, but it does not inform us about the specific reason for this priming effect in the standard AMP. Here is one example for an inference account for this finding: those who have shown priming in the previous task (the standard AMP) or in the first few trials of the IA-AMP might be more likely to infer that priming sometimes occurs, based on their observation of the congruency between their evaluation of the primes and their evaluation of the targets. Then, in the subsequent IA-AMP, when they see primes that elicit strong reactions from them, they would be more likely to predict that priming is about to happen. Is this an unconvincing post-hoc account? Perhaps. My point is that the present manuscript seems to suggest that new findings rule out a post-hoc inference account. Yet, I have not seen clear and coherent explanation why the authors think that this is the case. In contrast, it is clear that the novel findings are likely to inspire further research on specific accounts that could explain the priming effect in the standard AMP and the authors' present findings. In my view, this is why these experiments are important and could have a very positive effect on the scientific community.

**Authors**: see our previous comment. Once again, we are not “ruling out a post-hoc inferential account”. There can and will always be new post-hoc explanations offered once researchers inspect outcomes. We also agree that “novel findings are likely to inspire further research on specific accounts that could explain the priming effect in the standard AMP and the authors' present findings”.

With this in mind, we have included a new section in the General Discussion on alternative theoretical explanations for our findings, with a particular focus on the idea forwarded by the reviewer (see p.58-59).

**Reviewer 1**: 3. To continue the previous point, what did the authors mean, in p. 56, when they wrote that their results indicate the AMP effects "rely heavily on awareness of prime influence"? It might rely on people who show awareness, but that is not what the authors wrote. So, how does awareness lead to the priming effect? The authors seem to be careful not to talk about intention, perhaps because they do not have evidence about intention, but do they mean something else? Such statements left me sure about what the authors think their results suggest.

**Authors**: When we wrote that AMP effects “rely heavily on awareness of prime influence” we were referring to the repeated finding that (a) the AMP effect and its predictive validity appear to be based primarily on influence aware responding, (b) influence awareness rates vary widely between individuals but are highly consistent within individuals, within and between attitude domains, (c) participants who are more highly influence aware are responsible for group-level AMP effects, and that (d) recent modifications to the AMP that purportedly control for such subsample effects do not reduce or resolve this issue. Although non-influence aware trials retain some degree of predictive validity and contribute to some extent to the magnitude of effects, their contributions pale in comparison to that of influence aware trials.

That said, to avoid potential confusion, we have revised this sentence to now state “rely heavily on those participants who show awareness of the prime and its influence on their evaluations” (see p.56).

**Reviewer 1**: 4. Further continuing the previous points, when referring to the findings of Experiments 7 and 8, the authors wrote "In both studies the same pattern of findings emerged as before, findings that cannot be explained by a post-hoc confabulation account (given that there was nothing to confabulate)." In the introduction (p. 6), the concept "post-hoc confabulation account" was used to refer to the following hypothesis, attribute to Payne et al. (2013): "people may be able to identify that they acted in a particular way, but they are unable to say why they acted in this way (i.e., the post-hoc confabulation explanation)." I am not sure why this hypothesis is an account, and how it is related to the concept confabulation. Confabulation of what? Payne et al. (2013) argued that the reported intentional rating of the primes in Bar-Anan and Nosek's (2012) studies was confabulated. Did the authors mean confabulation of intention? Indeed, when trying to explain the results of Experiment 8, it is difficult to argue that participants wrongly think, before seeing the target stimulus, that priming would occur because they would intentionally rate the prime, while, in fact, priming is about to occur because of misattribution. Is that what the authors meant? Probably not because the authors wrote (p. 56) that the present work is agnostic about intention.

Alternatively, perhaps the authors use the term confabulation to distinguish between retrospectively inferred awareness after the priming had occurred (which one might consider confabulated awareness), and awareness that occurs before and during the priming effect. As before, this uncertainty about the authors' meaning would be solved if the authors explain better what they conclude from the results. Note, however, that people might be able to predict that an effect would occur, even if they are unaware of the reason for that effect. I might know that I am about to be sad today because I am always sad on Mondays, even if I attribute the effect of Mondays on my mood to the wrong reason. That is, people can confabulate the wrong reasons for a behavior that they predict.

**Authors**: After reading Reviewer 1’s comments we realized that the inclusion of post-hoc confabulation was leading to more confusion than clarity. The concept was also not central to the core aims of our paper. We therefore decided to omit discussion of this concept from the revised paper, and instead focus solely on the concept of awareness and the implications our findings have for theory and findings in the AMP literature. Therefore the above issue is no longer applicable to the revised manuscript.

**Reviewer 1**: 5. The authors found evidence for some awareness of the priming effect. But, I have not seen clear indication about how much awareness they found. For example, in Experiment 5 (pp. 35-36), the authors reported that the priming effect in IA-AMP trials that were flagged by the participants as sensitive to priming was superior in discriminating between Democrats and Republicans (d = 2.08) than the priming effect computed from the rest of the trials (d = 0.62). But, 0.62 is not a small effect, and the authors did not include a control group with a standard AMP that could provide an estimate for the overall effect in the AMP. Actually, I think the authors did not report the overall discrimination effect in the IA-AMP, to provide information about how much the "awareness" trials contributed to the overall discrimination effect. For example, if the overall effect, when including all trials, was d = 0.92, could we still conclude that the priming effect in the AMP relies mostly on "awareness" trials? I am not sure, but this information, and a more explicit discussion about the justifications for the authors' conclusion from that information would be helpful.

**Authors**: Based on the reviewer’s suggestions, we have included an estimate of the overall IA-AMP effect in Experiment 5 to provide information about how much the awareness trials contributed to the overall discrimination effect. We also note that that the confidence intervals of these three effects do not overlap, with influence-only AMP effects better at discriminating between individuals than an overall effect, which in turn was superior to a non-influence aware effect (see p.36).

**Reviewer 1**: It also would have been helpful to know more about the distribution of the frequency of the "awareness" trials. How many participants hardly reported about a priming effect? What was the typical report? Figure 1 of the supplementary materials is a good start but it aggregates across many experiments, and it does not provide clear information about the frequencies (notice also that, at least in my computer, the legend of that figure was unclear, and I did not understand how to distinguish between the two distributions displayed in that figure).

**Authors**: In line with the reviewer’s suggestion we have added a distribution of influence-awareness rates in the valence IA-AMPs in Experiments 2-8 (see Supplementary Materials).

**Reviewer 1**: Similarly, when the authors indicate that, within participants, priming in each trial was predicted by reported priming in each trial, one could still wonder whether that effect was driven by a minority of the participants. In other words, can we tell how valid that finding was, when using a similar logic to the logic that the authors used to question the validity of the priming effect itself? I hope the authors would be able to provide a better depiction of their findings that goes beyond the result in the statistical tests, by using more graphs, tables, and other reporting methods.

All that information becomes quite important for evaluating the authors' arguments about the practical implications from their results regarding the validity of the AMP. I was not sure that the authors provided strong evidence that "AMP effects are a poor index of 'general' evaluations in groups of people and a good measure of evaluations in highly influence aware people (who make up a minority of individuals in the task)." (p. 58). I do not remember seeing evidence about how small that minority was, or receiving any tools for estimating what minority size would be acceptable. To be clear, I believe that the authors' argument is valid, but not because of the (insufficient) information they provided about the present findings. Rather, I base my belief on the results reported in Bar-Anan & Nosek (2012) and our comparison between the AMP and other indirect measures, that I mentioned in my previous review (Appendix D of Bar-Anan & Nosek, 2014, at <https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-013-0410-6/MediaObjects/13428_2013_410_MOESM1_ESM.pdf>).

**Authors**:

**Reviewer 1**: 6. It would help the readers if the authors explicitly mention and discuss the fact that in Experiment 1, unlike in the experiment it attempted to replicate, participants completed an AMP before completing the AMP that allowed skipping. In other words, unlike the original experiment, in the present replication, there was a confound between the AMP's type (skip or standard) and whether participants have previously completed an AMP.

**Authors**: We have added the requested information into the revised manuscript (see footnote 5 on p.15). Specifically we now state:

“As we previously noted, we adopted a within participant design (participants first completed a standard AMP followed by a ‘skip’ AMP) whereas Payne et al. (2013) adopted a between participant design (participants completed either a standard AMP or a ‘skip’ AMP). Thus, unlike the original experiment, performance on the ‘skip’ AMP was always influenced by prior completion of a standard AMP.”

**Reviewer 1**: I should note that despite this confound, I do believe that the replication's results might generalize beyond the particular setup of their experiment. In 2008, together with Keith Payne (who originally collaborated with Nosek and me), we have conducted a replication of the skip-AMP that was never published. Like the authors of this manuscript, we found that participants in the no-skipping group showed slightly more extreme AMP attitude (M = .17, SD = .16) than participants in the skip-option group (M = .15, SD = .12), t(1,211) = 2.97, p = .003, d = .14, indicating that the option to skip slightly helped in decreasing the priming effect. However, I vaguely remember that our results depended on specific rules for exclusion of participants. Therefore, without pre-registration I did not consider the small effect from our 2008 study a robust effect. The new replication increases my belief in the possibility that Payne et al.'s (2013) original finding is not easily generalized to other contexts and samples.

**Authors**: We thank Reviewer 1 for sharing this information with us.

**Reviewer 1**: 7. The authors wrote in p. 56 "On the one hand, AMP effects may reflect misattribution, as is often claimed, yet people are fully aware that misattribution is taking place". As I noted earlier, the finding that participants are aware of the priming effect is not evidence that they are aware of how it occurred (e.g., misattribution).

**Authors**: We have included new material highlighting this possibility in the General Discussion (see p.56-58).

**Reviewer 1**: 8. In p. 6, the authors mischaracterized Bar-Anan and Nosek's (2012) findings: "They found that AMP effects were larger, more reliable, and primarily moderated by those who did so (i.e., intentionally rated the prime rather than the targets)." The AMP effects were moderated by those who retrospectively reported intentionally rating the primes. We explained in our paper that we could not know, from our studies, whether this report was accurate or confabulated.

**Authors**: we have revised our treatment of Bar-Anan and Nosek (2012) based on the reviewer’s comments (see p.6). Specifically, we now state:

“For instance, Bar-Anan and Nosek (2012) asked participants to first complete an AMP and then retrospectively indicate if they had intentionally based their evaluations on the prime rather the target. They found that AMP effects were moderated by those who retrospectively reported intentionally rating the primes.”  
  
**Reviewer 1**: 9. In p. 7, the authors wrote that the accounts for the AMP effect "differ in the role that awareness is assumed to play in AMP effects, with proponents of the implicit account arguing that the prime stimuli influence participants' evaluations without their awareness, while proponents of the explicit account argue that participants are aware of the influence of the primes on their responses." Notice, however, that arguments about whether people are aware of the priming effect are not arguments about the role of awareness in the AMP effect. I think that "role" refers to the causal nature of awareness.

**Authors**: we have revised this section in line with reviewer 1’s request (see p.7). Specifically, we now state:

“The implicit and explicit accounts also differ in how awareness and AMP effects are thought to be related to one another. Although both acknowledge that people can be aware that the prime has influenced their response to the target, they differ in the causal vs. correlational role that awareness is assumed to play. Proponents of the implicit account argue that awareness may be correlated with, but is not causally required to demonstrate, AMP effects, whereas proponents of the explicit account argue that participants are aware of the influence of the primes on their responses, and that it is possible this awareness is causally related to those effects.”  
  
**Reviewer 1**: 10. In p.59, the authors wrote "Yet our findings suggest that the neutral AMP effect observed in this officer does not mean that the officer has no particular racial evaluations. It may be the case that the officer holds very strong anti-black evaluations but does not produce an AMP effect due to his low influence awareness rate." I do not think that the authors showed that there were many (or any) participants who have extreme attitudes but these are not shown in the AMP, or that this is more likely in the AMP than in any other psychological measure. Further, the authors seem to argue that their findings suggest that self-reported priming is the reason for the priming (no AMP effect due to low influence awareness rate). I do not think that the authors found evidence about such a causal effect, and it is not clear how that would occur - how does awareness cause the priming effect in the AMP? This question might circle back to my earlier comments in this review.

**Authors**: This section of the paper was removed during the most recent round of reviews. As such, this comment no longer applies to the current version of the manuscript.  
  
**Reviewer 1**: 11. Throughout the article, the authors often use the term "influence awareness" when they refer to participants' reports that priming has occurred. Reporting a priming effect is different than awareness of the priming effect. If priming effect had not occurred and I report that it occurred, it would be wrong to argue that I am aware of the priming effect. The correlation between the self-reported priming effect and the priming effect suggests that, in general, some people showed some awareness of the priming effect.  
  
For example, the authors used the term "influence awareness rates" to refer to the rate of trials in which the participants reported that the prime influenced the evaluation of the target. We do not know that in each of those trials the prime indeed influenced the evaluation of the target. Therefore, it is inaccurate to use the word "awareness" to describe this rate.

**Authors**: We see the reviewers point here (reporting a priming effect is different than awareness of the priming effect). For instance, imagine a situation where no priming effect is present. If a participant says that the prime influenced their response in this situation then it is not an accurate awareness of the prime’s influence on their target responses. If anything it is an *inaccurate* self-reported awareness of prime influence.

Now imagine that a priming effect is present. If the participant says the prime influence their evaluations and it did so then this would be an *accurate* awareness of prime influence. Put simply, there is a 2 x 2 awareness /accuracy matrix that is possible here:

Reported Awareness: Yes No

Accuracy of that Report: Yes No

Now in our studies we examined, for each individual trial, if saying one’s evaluations were (Experiments 2-6) or will be (Experiments 7-8) influenced by the prime predicted more prime-consistent than prime-inconsistent responding, compared to those who said they were not influenced by the prime. Thus when we refer to “influence awareness” rate we are referring to the rate at which Reported Awareness (YES) differed from Reported Awareness (No) in terms of the Accuracy of that Report (YES).

In the revised manuscript we have included a table that outlines the number of times that consistent and inconsistent prime-evaluation responses co-occur with trials that were registered as influence-aware and non-influence-aware. That is, the number of times that Reported Awareness (Yes or No) co-occurs with the Accuracy of that Report (Yes or No) in each of our experiments (see Table 3 in the Supplementary Materials).

Reviewer #2: Review: Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness  
  
First, I should start by saying I was not involved in the review process since the beginning, and thus I have no basis for evaluating whether the manuscript improved relative to the first submission. Moreover, I should also clarify that I didn't read the previous reviews + authors' responses before submitting this review.  
  
The current manuscript presents eight experiments examining the role of awareness in driving AMP effects. Taken together, the results of these different experiments don't leave any doubt that influence awareness plays a huge role in AMP effects, which is indeed inconsistent with an explanation for AMP effects based on misattribution. In my view, these findings represent an important contribution to the vast literature on implicit social cognition, particularly for studies using the AMP, and may also have implications for other research domains. For example, several studies on the relationship between positivity and familiarity and fluency and familiarity (for a review, see Winkielman et al., 2003) draw on misattribution to explain its findings.

**Authors**: we thank the reviewer for their kind words.  
  
**Reviewer 2**: Although I consider this to be important work, there are some minor issues that I believe need further explanation.  
  
1. Previous studies suggest that when participants are aware of the influence of primes on judgments, they discount that influence, thus reducing or eliminating misattribution effects (e.g., Gellatly et al., 1995; Jones et al., 2009; Oikawa et al., 2011; Ruys et al., 2012; White & Knight, 1984). Interestingly, in the present experiments, aware participants seem to be doing the opposite: instead of discounting the primes, they deliberately use the primes to judge the targets. Is this true? In other words, what is the % of aware trials in which participants responded accordingly to the prime (e.g., positive prime = positive judgment) vs. not accordingly (e.g., negative prime = positive judgment)?

**Authors**: See our response to the final comment of Reviewer 1. We have now included a table outlining the rate of prime-consistent and inconsistent responses when people report that the prime influenced their responses vs. reported that the prime did not influence their responses (see Table 3 in Supplementary Materials).  
  
**Reviewer 2**: 2. The authors refer to Bar-Anan and Nosek (2016) as important research showing that awareness plays a role in the AMP; however, it is not entirely clear in what aspects the present work differs from Bar-Anan & Nosek.

**Authors**: during the latest revisions to the manuscript this material was removed. Thus this point is no longer applicable.

**Reviewer 2**: 3. Given that the goal of the present research is to examine the role of awareness in AMP effects, it would have been relevant to control whether participants performed the task before or not. Did the authors ask participants whether they were familiarized with the AMP?

**Authors**: We did not. But we agree that this is a nice point and have added material in the General Discussion that speaks to the issue (see footnote 12 on p.63). Specifically, we now state:

“One could also explore other factors that may impact how much one is likely to be influence aware during an AMP, from one’s previous experience with the task (e.g., extensive vs. limited prior AMP exposure), as well as their motivation or opportunity to process information during the task in an effortful manner.”

**Reviewer 2**: 4. In experiment 1, the authors obtained a larger AMP effect in the standard AMP than the skip-AMP. Can this result be explained by a statistical power difference between the tasks, given that AMP scores were based only on the non-skipped trials in the skip-AMP vs. all trials in the standard AMP?

**Authors**:   
  
**Reviewer 2**: 5. What were the primes employed in the political AMPs used in experiments 4 and 5? IAPS pictures?

**Authors**: the prime stimuli used in these experiments consisted of six images of Donald Trump and six images of Barack Obama taken from the Presidents-IAT materials of the Project Implicit website (see osf.io/f38ag).  
   
**Reviewer 2**: 6. I think it is confusing to report between-subjects Cohen's d.

**Authors**: It is not clear what actually the reviewer is referring to here given that (a) we report many Cohen’s d’s throughout our paper, and (b) Cohen’s d’s by definition involve between-subjects comparisons (within-subjects involves Cohen’s dz).   
  
**Reviewer 2**: 7. In the GD, the authors propose that future research should try to understand what makes a person influence aware. I agree that this is an interesting future avenue for research, but I also wonder why some primes are more influence-aware than others? Are there any common features among the most influence-aware primes (e.g., valence extremity)?

**Authors**: we agree with the reviewer that this is also an interesting question. We have included additional material in the General Discussion which touches upon this topic (see footnote 12 on p.63). Specifically, we state the following:

“Likewise, one could examine what properties of the AMP alter one’s likelihood of being influence aware. Is it that certain types of primes that are more likely to elicit influence awareness than others (e.g., because of their valence, extremity, familiarity). Is it that certain types of instructions direct attention towards vs. away from the prime, or the relationship between the prime and target, and this also alters influence awareness? Both topics seems worthy of future investigation. ”  
  
**Reviewer 3**: This revised paper includes the previous studies plus three additional ones. The problems with the original studies remain as they were in the first submission. One new study (now study 1) used a skip method from Payne et al (2013) with a modified within-subjects design. Whereas the 2013 study found that participants showed significant affective priming effects when they had a skip option available, and when they did not. The new study 1 also found significant priming effects in both conditions, though unlike the original study, there was a significant moderating effect, with smaller priming effects when a skip option was available. In a second new study (now study 7) subjects saw the prime and target, and were asked to rate the expected influence of the prime on their target rating before expressing their rating. And in the final new study (now study 8) they saw a prime and then rated the expected influence of the prime before the target was presented and before they rated the target. In all of the studies, participants showed above-chance accuracy in their ratings of prime influence. That is, participants with larger priming effects reported greater perceived influence from the primes. The authors use reported influence ratings as a moderator of priming effects, and conclude that the AMP is only valid for a subset of participants and/or trials in which participants are aware of the primes' influence.

This revision removes much of the inappropriate causal language that was used in the original draft to describe correlational findings. It also has deleted many of the inaccuracies and mischaracterizations of previous work that were in the previous draft. However, it has replaced those with new inappropriate causal inferences and new mischaracterizations, as I describe in detail below.  
  
Study 1: Study 1 is informative, and suggests that participants have awareness of the primes' potential to influence them that is greater than zero, at least after they have experience with the task. The study is described as a "failure to replicate" the 2013, which is not accurate, both based on the design of the study and the logic of the analysis and conclusions. In the 2013 study we manipulated the skip option between subjects so that subjects could not observe their behavior in the non-skip task and use that self-perception to inform their skipping decisions in the skip version. The present study used a repeated measures design in which all subjects completed the non-skip version first, and then completed the skip version, maximizing the chance that subjects could base their skip decisions on observations of their previous behavior. This design likely overstates awareness compared to the between-subjects design. So this is not a direct replication, as the design differs in important ways. Moreover, a within-subjects design will have power to detect smaller effects than a between-subjects design, so an effect that is significant in the within, but not between-subjects design likely reflects design choices. It is strange to call this a failed replication.

**Authors**: We agree with Reviewer 3. We did not carry out a *direct* replication but rather a *conceptual* replication of Payne et al. (2013; Experiment 3). We have now revised the manuscript to acknowledge this. We have also included material acknowledging that the prior completion of a standard AMP could raise influence awareness rates on the skip AMP (see p.10, 16, 18).

**Reviewer 3**: More importantly, the paper misstates the logic of the study. The logic of the original was that if a person has \*perfect\* insight into when primes influence their judgments and can use that to regulate their behavior, then they would always know when to skip, and this would eliminate the priming effect (it would be zero). On the other hand, if a person had \*zero\* insight, then they would skip randomly and it would not reduce the priming effect at all. The present results fall in the middle, suggesting non-zero insight, but nowhere near perfect insight. There was apparently significant priming in both conditions. But the data are not reported in a way that shows the priming effect in each condition. The priming effect should be shown in each condition. The significant priming effect in the skip condition questions the authors' claim that the AMP is not valid among unaware trials/subjects. Why would primes influence target judgments even when people have the option to skip when they think they are influenced?

**Authors**: With respect to Reviewer 3’s first point, we cannot find any reference to \*perfect\* nor \*zero\* insight in Payne et al. (2013; Experiment 3). If anything the original authors argued that “By comparing the ***magnitude*** of priming in the two conditions, we can estimate ***how much awareness*** participants had about the influence of primes in real time as they formed their evaluations of the pictographs.” (p.382).

With regards to the second point (“there was apparently significant priming in both conditions” and “the data are not reported in a way that shows the priming effect in each condition”) we are alittle confused here. We never set out to report or examine this.

The original theoretical claim made by Payne et al. (2013) focused on the *difference* between the size of ‘skip’ and ‘standard’ AMP effects rather than their independent sizes. Nevertheless, and in-line with the reviewer’s request, we have included means and SDs for both AMPs in the revised manuscript (see p.18). When we did so it became apparent that the mean of the skip- AMP was ***opposite*** to the direction that was predicted based on the original Payne et al. study and the underlying theoretical assumptions.

**Reviewer 3**: Other than study 1, the reported studies all suffer from the same flaw as the original submission, which is a simple alternative explanation: Subjects infer the influence of the primes from the intensity of their affective reaction to the primes (or their attitude strength, for which the priming effect is a proxy in this task). The difference between explanations is critical for evaluating what claims can be supported by these data: (1) If the claim is simply that participants can make inferences about whether the prime influences their ratings of the targets that are more accurate than chance, then the data are consistent with that. (2) But if the claim is that people have insight into the causal impact of the primes on their ratings of targets, then the picture is much murkier. (3) And if the claim is that AMP effects are caused by a subset of participants who are aware of the causal influence of the primes on ratings of targets, then the studies offer no evidence for this claim, as detailed more below.  
  
Studies 2-6  
  
Although the authors have scrubbed some of the causal language from the paper, replacing "driven by" with "attributable to," "explained by," and so on, the claims made are still causal arguments, supported by correlational data. In the AMP tasks, subjects saw primes and targets, then rated the target, and then rated whether they were influenced. An analysis following this temporal order would use the prime congruence of target ratings as the independent variable, and rated influence as the dependent variable. An accurate and unbiased interpretation of that effect would be "when subjects responded to the target in a prime-congruent way, they were subsequently more likely to claim that they were influenced by the prime. We cannot say based on these correlational data whether prime-consistent responses caused higher ratings of influence, or whether the perception that they were influenced caused more prime-congruent target ratings. Or it might be that a third variable, such as the intensity of emotional reactions to the primes might explain both prime-congruent responses and ratings of influence."  
  
But the authors do not make that interpretation, and they do not report that analysis. Instead, they make ratings of the target the DV, and use rated influence as the moderator. Although they have removed the word "drive," they still draw a causal conclusion that awareness contributes to AMP priming effects, or that AMP priming effects depend on awareness. This is an elementary confusion of correlation and causation.

**Authors**: We politely disagree with the reviewer for several reasons. First, for the “retrospective awareness” studies the reviewer refers to, we repeatedly examined *predictive* (postdictive) relationships between influence awareness at Time 2 and (standard) AMP effect sizes at Time 1. We consistently found that higher influence-awareness rates *predicted* (postdicted) higher scores on a previously-completely standard AMP.

Second, the aforementioned temporal account is not applicable to Experiments 7-8 were awareness was measured prospectively rather than retrospectively. Participants either (a) saw the prime, target, made an awareness response, and then an evaluation, or (b) saw the prime, made an awareness judgement, then the target then the response. It seems entirely reasonable to us to treat ratings as a DV and influence awareness judgements as an IV here. Critically, these studies also produced near identical outcomes to those reported in our previous studies.

**Reviewer 3**: Studies 7 & 8  
  
These studies measure perceived prime influence before subjects express their target rating (study 7) or before they see the target (study 8). These studies show, even more clearly than studies 2-6 that the observed pattern is consistent with the explanation that subjects make an inference from their affective reactions to the primes to how much their judgments will be influenced by it. In these designs, they must be doing so because there's no actual target judgment yet to have insight about. So in these designs, any confabulation is not "post-hoc" about the target rating, but it would still be a confabulation based on affective experience in response to the prime. So, if the claim is simply that people can accurately guess how much primes influence them in the AMP, this conclusion is supported by the data, but it has no bearing on the question of how the AMP works or its validity, as I expand on next.

**Authors**: We acknowledge that reviewer’s idea is possible. However, it seems that this idea involves a causal inference being drawn from correlational data – the very type of inference that was previously argued to be problematic (e.g., “…any confabulation is not "post-hoc" about the target rating, but it would still be a confabulation based on affective experience in response to the prime”).

As we noted in our response to Reviewer 1, such an account also seems to be post-hoc and conditioned on the observation of the outcomes. It is not an *a priori* prediction nor was it pre-registered (unlike those forwarded in this paper). This is especially true for those theoretical claims made for Experiments 7-8. Reviewer 1 acknowledges this when he says “Payne et al. (2013) suggested that the awareness of the priming effect is only retrospective, inferred after the fact, rather than something that exists while the priming occurs.”

In short, we recognize our interpretation of the data differs from the reviewer’s. We also believe this represents a healthy scholarly debate and highlights one of the ways in which our paper can contribute to the field’s use and understanding of the AMP. In the spirit of intellectual fairness, we have included new material outlining the reviewer’s ideas in the General Discussion, and recognize them as an alternative (post-hoc) account of our findings (see p.58-60).

**Reviewer 3**: Conclusions for validity of the AMP

Regarding the putative misattribution mechanism, the paper says that if subjects know that the prime is influencing their target ratings, it can't be a misattribution, but I don't agree. A misattribution, by its nature of being a mistake, can't be entirely understood by the person making the misattribution, otherwise they would presumably not make that error. But there are many ways to have mistaken or incomplete knowledge about the sources of one's thoughts and behavior. For example, I might be aware that standing on a suspension bridge is making me nervous, which might influence my attraction to another person standing on the bridge. And yet, I might still think the person is attractive. If the bridge caused that perception, then it is still a misattribution. In the AMP, subjects might accurately guess that their evaluations of targets are influenced by primes (especially when they feel strongly about the primes) but also believe that the target pictograph really is a very pleasant pictograph. That would still be a misattribution, and is entirely consistent with these studies.  
  
There is a large literature about the ways that people use lay theories to predict or explain their own behavior, and why those inferences do not necessarily reveal insight into people's cognitive processes (see Nisbett & Wilson, 1977; Wilson, Hull, & Johnson, 1981; Wilson & Dunn, 2004). And there is a related literature on authorship processing and the experience of conscious will showing that people draw inferences about the causes of mental processes that do not necessarily track true causes (Aarts et al., 2005; Kühn and Brass, 2009; Wegner & Wheatley, 1999). These are deep problems in the attribution of mental processes that are not easily solved by simply asking people how or why they were influenced. But this paper neglects these entire fields, simply asks people if they were influenced, and then takes those judgments at face value as accurate causes of mental processes.

**Authors**: We now include new material in the General Discussion that incorporates the reviewer’s ideas on misattribution (see p.56-58). Specifically we now say:

“On the one hand, AMP effects may reflect misattribution, as is often claimed, yet people are fully aware that misattribution is taking place. Moreover, our findings with prospective measures in Experiments 7-8 would require people to not only be aware of misattribution but also be able to predict that it is going to occur even before a target is evaluated or a target stimulus is even presented. Although this idea runs contrary to how misattribution is traditionally defined (Schwarz & Clore, 1983) it is possible. For instance, one reviewer of this paper argued that misattributions are - by definition - mistakes. As such, they cannot be entirely understood by the person making the misattribution otherwise they would presumably not make that error. When it comes to the AMP participants may be aware that their target evaluations are influenced by the primes (especially in cases where they feel strongly about the primes). But they may also hold the mistaken believe that the target really does have a particular valence. This latter belief would still qualify as a misattribution. Yet even in this case, our findings suggest that misattribution would still be occurring or captured in only those participants who were highly influence aware, rather than people in general. As such, changing the conceptualization of misattribution does not by itself address the issues raised by our findings.”

**Reviewer 3**: The most serious error in the paper is to use reported influence as a moderator, and then conclude that AMP effects are valid only for the subjects who reported influence. This is the same problem I wrote about in the first review, of taking an outcome variable and using it as a predictor or a moderator, known as "post-treatment bias" or post-treatment conditioning (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.  
  
Concretely, if the AMP is equally valid for all subjects and judgments of influence are caused by attitude strength (i.e., the strength of affective reactions to the primes) then dividing people into "more aware" and "less aware" based on the influence ratings is guaranteed to find the observed results. That is because subjects with stronger attitudes will report more influence of the primes as a consequence. Then when subjects are separated by rated influence, the ones who report little influence will be the subjects with weak attitudes and little variability on AMP scores. The group with high rated influence, in contrast, will have strong attitudes and more extreme priming effects. Again, this is guaranteed to happen even if the AMP is equally valid for all subjects and awareness plays no role in driving the priming effect.

**Authors**: The Reviewer’s argument rests on the fundamental assumption that “the AMP is equally valid for all subjects”. Yet our data strongly suggests otherwise. We consistently find that it is only a *subset* *of influence aware trials* that drive the effect within-individuals. And between-individuals, it is only a *subset of individuals* that drive group level effects. Thus the premise upon which the conclusion is based does not seem to be in line with the data being reported here (i.e., the AMP is not equally valid for all subjects).

With respective to the idea that “judgements of influence are caused by attitude strength”, this is again a causal inference made on the basis of correlational data (the very same issue the author had with our own theorizing). That said, in the spirit of intellectual fairness, we have included new material in the General Discussion that speaks to the reviewer’s claims, and acknowledge them as one of several alternative explanations for our findings (see p.58-60). Specifically we state the following:

“Another reviewer advanced an attitude strength perspective on our findings. The idea here is that if the AMP is equally valid for all participants, and judgments of influence are caused by attitude strength (i.e., the strength of affective reactions to the primes), then participants with stronger attitudes will report more influence of the primes as a consequence. When participants are separated based on their influence awareness ratings, the ones who report little influence will be those with weak attitudes and little variability on AMP scores. Those who report high influence will have strong attitudes and more extreme priming effects. Thus influence awareness is a by-product or corollary of attitude strength rather than the main factor driving AMP effects.”

**Reviewer 3**: The authors argue that the same logic about validity would apply to other tests such as the IAT. I agree that the issues are the same, but the authors' analysis is equally misleading for the IAT. Research suggests that people can indeed report with some accuracy about the size of their congruity effect on the IAT (See Adam Hahn and Bertram Gawronski's work). If you ask who was influenced by congruity in the IAT and then use influence ratings as a moderator, will find that the effect is larger for those who report more influence, and more strongly correlated with other variables. But this is not because the task is only valid for those who are aware of it; it is because we have conditioned on a variable that is downstream from the congruity effect.

**Authors**: See our previous comments on temporal ordering and treatment of influence awareness as an IV and DV.  
  
**Reviewer 3**. To summarize, the present paper was not responsive to the previous reviews. With the exception of study 1, the present studies do not address the problems identified in the first submission, and they do not make a substantial advance beyond the Bar-Anan / Payne exchange in 2012/2013. Although the revision removed certain causal words, the causal assumptions, analyses, and interpretations remain, and the elementary confusion between correlation and causation remains.

**Authors**: We find it unfortunate that the reviewer feels we were “not responsive to the previous reviews”. This is inconsistent with what Reviewer 1 feels and our own perspective. Across multiple rounds of reviews we have genuinely engaged with his comments. We added three new studies (one replication and two novel empirical studies) that speak to the theoretical issues he previously raised (retrospective vs. prospective awareness). We have significantly revised the paper multiple times to correct for inaccuracies. We now include several additional sections that highlight his theoretical ideas as one of several alternative accounts of our findings.

As we note in one of our responses to Reviewer 1, our work does go beyond the Bar-Anan / Payne exchange by investigating awareness in both retrospective and prospective ways (something that was not done before). We do so using multiple measures (single post-hoc self-reports, trial-by-trial online measures; something that was not done before), versions of the AMP (standard, Influence Aware version, Mann et al. version; something that was not done before), and attitude domains (political, positive vs. negative; something that was not done before). They explore the bidirectional (predictive) relationship between performance on one AMP and that of another (something that has not been examined in prior awareness work), and examine how awareness plays out at the trial, individual, and group levels (something that was not done before). The result is a level of generalizability about awareness in the AMP that extends beyond that which has come before.