**Editor**: Thank you for submitting your research to Behavior Research Methods. Our referees have now considered your paper (see their comments at the bottom of this letter). As you can see, the reviewers find your submission worthwhile but would like to see minor to medium revisions made to your manuscript before it can be published. This agrees with my own reading. Therefore, I invite you to revise your manuscript. Please note that my invitation is no guarantee of eventual acceptance. (My own feelings would be pro-mixed models, but I can see the analysis going either way).

**Authors**: We thank both the Editors and Reviewers 1-2 for their constructive and kind comments. Based on this feedback, we have revised the paper. Specifically, in line with Reviewer 1’s request, we have included additional content discussing what might be happening during a typical AMP trial, as well as the need for greater precision on concepts such as (un)awareness and intentionality. We also provide more methodological detail in the manuscript itself as well as simple correlations and descriptive statistics. With respect to Reviewer 2, we respond to the use (and appropriateness) of mixed models in our paper, as well as the request to include descriptive statistics. Overall, these suggestions have further improved our paper and we thank both Reviewer’s for their time and effort.

**Reviewer 1**: Thank you for the opportunity to review this paper. The research question is interesting and important, and the justification for the work was laid out clearly (with one exception, *see below*). Overall, I’m highly enthusiastic about this manuscript.

**Authors**: We thank Reviewer 1 for her kind words and constructive comments.

**Reviewer 1**: There are two (maybe three) substantial issues that I would like to see addressed more thoroughly, but I recognize that no paper can exhaustively address all concerns and see merit in the paper being published despite these limitations - if they can at least be discussed in greater detail.

First, despite the authors’ strong effort, I am not convinced that “typical” performance on an AMP is necessarily comparable to one where they stop and respond to questions after each trial. There’s the possibility that responding to the items may make them self-conscious about their performance, and attune them to possibilities (e.g., rating the prime) that they might not have spontaneously attempted on their own, and also the possibility that stopping after each trial reduces a “flow” state where participants begin to respond quickly, habitually, and with a gut reaction. I appreciate the attempts to address this concern, but I’m not convinced it was entirely successful.

**Authors**: We appreciate the Reviewer’s point here. It may be that during a standard AMP, participants don’t necessarily pause after each and every prime-target presentation, reflect on what they just saw, and then use this information to inform every individual target evaluation that they make. Nevertheless, our work does show that there are *deep* connections between what is happening in the IA-AMP and what is happening in a standard AMP. As we repeatedly show, performance on the IA-AMP predicts the magnitude of standard AMPs that either come before or after it, both within and between different attitude domains. This occurs when the standard AMP is modified to reduce or resolve this issue, and when influence awareness is measured in different prospective and retrospective manners.

We can see multiple possibilities that could explain this relationship between IA-AMP and standard AMP. It may be that participants don’t stop and reflect on every trial during a standard AMP. But it could be that they stop and reflect on *multiple* trials, and this awareness of the prime’s influence grows as one completes the task. Alternatively, it may be that reflecting on a single or few trial(s) early on in the task may lead them to generate inferences that influence their subsequent performance on the task (e.g., “Certain primes keep influencing how I respond…” or “Certain stimuli [primes] keep flashing up on screen…maybe I’m supposed to respond to them in a particular way…”). In other words, as the Reviewer suggests, it’s certainly possible that responding to the targets than follow primes may make people conscious about their performance, and attune them to possible explanations for that performance, a process that influences how they subsequently respond.

In-line with the Reviewer’s suggestions, we have added new content to the General Discussion that elaborates on the various possibilities that may help explain why there is a deep connection between performance on the IA-AMP and standard AMP, and what might be happening during the standard AMP itself (see changes on p.68 in **blue**).

“ *Our findings highlight a deep connection between performance on the IA-AMP and standard AMP. The fact that performance on the IA-AMP predicts the magnitude of standard AMPs that either come before or after it, both within and between different attitude domains, and when influence awareness is measured in different prospective and retrospective manners indicates that what holds in the former likely plays out in the latter. But this raises the question – how does influence awareness play out in standard AMPs? Is it that participants pause after each and every prime-target presentation, reflect on what they just saw, and then use this information to inform every individual target evaluation that they make? This seems unlikely to us given that trial latencies (while not critical to the effect or the task) are typically short. One alternative is that people don’t stop and reflect on every standard AMP trial but do stop and reflect on multiple trials, and this awareness of the prime’s influence grows as one completes the task. Yet another alternative is that stopping and reflecting on a single trial early on in the task may lead people to generate inferences that influence how they subsequently perform (e.g., “Certain primes keep influencing how I respond…” or “Certain stimuli [primes] keep flashing up on screen…maybe I’m supposed to respond to them in a particular way…”). In other words, as one Reviewer suggested, responding to the targets than follow primes may make people self-conscious about their performance, and attune them to possible explanations for that performance, that subsequently influence it. In short, future research could examine how influence awareness actually plays out within the standard AMP. This would not only provide useful information about the task itself, but also about influence awareness as a process.*”

**Reviewer 1**: My second major concern has two (related) parts. To start, too much methodological detail was relegated to supplementary materials. I had to really go searching to figure out what participants actually responded to in these tasks. This hunt is even more difficult in the latter studies where the methods say something like measures are “identical to Study 2” but then you go back to Study 2 and find a directive to go to the supplement. Then, when I got to the supplement, it was really difficult to figure out what is going on.

**Authors**: In line with Reviewer 1’s suggestions, we have now revised the manuscript to provide better clarity around our various methods and designs (see changes to the various Procedure sections throughout the manuscript in **blue**).

**Reviewer 1**: It also seems like the analysis deviates from the pre-reg (or maybe the study itself does?). I found a pre-registration document that includes four items: *Influence awareness*: On how many trials was a valenced picture presented before the Chinese character? It is important that you are honest here.” *General influence*: To what extent were your ratings of the Chinese symbols influenced by the pictures that appeared immediately before those symbols?” *Intentional influence*: How often did you intentionally base your rating of the Chinese symbol on the image that immediately appeared before it? *Unintentional influence*: How often do you think that your rating of the Chinese symbol was unintentionally influenced by the pictures that appeared immediately before those symbols? The manuscript says that the “general influence” item is the only one pre-registered and the others are exploratory, but I am not seeing that anywhere in the actual pre-registration document.

**Authors**: we thank the Reviewer for highlighting this for us. And our apologies for any confusion here. This is entirely to do with our lack of precision in the pre-registration document.

First, in line with the previous comment, we make it clear in the procedure sections that self-report measures were taken at the end of our studies. In many cases these assessed for influence awareness, general influence, intentional influence, and unintentional influence.

Second, in Experiment 2, we pre-registered an interest in analyzing one of these self-reported measures. Confusingly, we referred to this as “General influence” in the pre-registration document and as an “offline measure of influence” in the analysis section of that same document (see Hypothesis 3 in the pre-reg document for Experiment 2).

Put simply, we did pre-register an interest in whether the post-hoc self-reported awareness measure (i.e., what we labelled as “General Influence”) correlated with the trial-by-trial (online) measure of influence awareness we obtained from the IA-AMP. But we did a bad job at articulating this in the pre-registration document. We have added a short footnote to the manuscript to clarify this point for readers (see footnote 7 on p.22 in **blue**).

“*Note that in the interest of clarity we referred to this self-reported measure of awareness as ‘General Influence’ in our pre-registration document and as an offline measure of awareness in the analytic section of that same document*.”

**Reviewer 1**: To be clear, I am not trying for some kind of “gotcha”! It wasn’t my intention to compare the method to the pre-reg, and it’s possible that I missed something when I did. My reason for digging around to figure out what items participants actually responded to (which I still do not fully understand), is that “awareness” and “intentionality” are entirely muddled, both conceptually and methodologically, throughout this paper. The authors seem to prefer the interpretation that the “to what extent where your ratings influenced” item gets at awareness, which it might, but this is not a clean interpretation because the questions do not seem to be asked in a way that allows disambiguation (or they are but the full results for those items that could help with interpretation are not presented to us).

**Authors**: We agree with Reviewer 1. “Awareness” and “intentionality” are not clearly delineated concepts and this is a systemic and ongoing issue for the field of implicit cognition as a whole: there is massive variation in what researchers refer to when they use these (and other) terms related to automaticity and implicitness (for an excellent treatment of conceptual imprecision in this area we highly recommend Corneille & Hütter, 2020; *PSPR*).

Imagine, for instance, that a researcher is interested in implicit racial bias. There are many conceptual degrees of freedom available to them. They could define “awareness” as a participant’s ability to self-report such a bias (i.e., that they have an automatic evaluative preference for one group over another). They could use the very same term to refer to a participant’s ability to report on *how* that automatic response bias is being assessed (e.g., via speeded categorization of one social group with positive words and a second group with negative words as in the IAT). Or the presence of particular stimuli and response requirements in a given procedure (e.g., that a black person is being presented as a prime stimulus). Or the capacity of the participant to identify the *origin* of this implicit bias (e.g., as stemming from society or their personal learning history).

The very same applies to the other terms typically subsumed under the umbrella of “automaticity” or “implicit”, whether it is “intention”, “speed”, “efficient” or “unconscious”. This is to say nothing of the remarkable differences in how these various terms have been measured in the wider implicit literature (e.g., awareness is often measured using on- or offline measures, in retrospective or prospective manners). Reviewer 1 nicely acknowledges this in their reference to Hahn and Goedderz (2020).

With the above in mind, we avoided using general and ambiguous terms like awareness and intention in our work, and instead tried our best to precisely operationalize what it was that we were studying (i.e., *influence* awareness; the capacity for a person to self-report that the prime has influenced how they have responded towards a target stimulus). We also purposefully measured this in multiple ways (trial-by-trial vs. post-hoc; prospective vs. retrospective) to ensure that our arguments were not constrained to one methodology or approach. As we report below, we have also added new material clarifying precisely what we did and did not measure, and how other automaticity conditions related to the AMP could be studied in future work.

**Reviewer 1**: To me it seems that there’s an important difference between “yes I am aware I was evaluating the stimulus based on the prime because I was doing it on purpose” and “yes I am aware I was evaluating the stimulus based on the prime and I couldn’t stop myself even though I tried”. This is an important distinction because intentionality/automaticity are part of the definition of “implicitness” and a better understanding of these concepts can help us to better understand on what dimensions we might think of the AMP as being implicit (or even if whether or not it’s an indirect measure at all – if people are intentionally evaluating the stimulus it may not be an indirect measure either). Couldn’t the title of the paper just as easily be “AMP Effects are Strongly Moderated by Intentional Responding”?

**Authors**: As we note in footnote 1, for us, the term ‘implicit’ does not represent an all-or-nothing concept, but rather is an umbrella term which refers to a set of automaticity conditions under which mental processes are said to operate. The effects obtained from an indirect procedure are assumed to occur under one or more of these automaticity conditions. Thus to describe a measure or effect as implicit requires that one is clear about the exact automaticity conditions relevant to that effect.

In the current paper we make it clear that our focus is on *one* particular automaticity condition (awareness), and one particular type of awareness (influence awareness).

What the Reviewer seems to be proposing here is a question about the interaction/relationship between automaticity conditions. We agree that there may be an important difference between responding in an intentional *and* aware manner (“*I am aware I was evaluating the stimulus based on the prime AND I was doing this on purpose*”) or aware *and* unintentional manner (“*I am aware I was evaluating the stimulus based on the prime AND I was NOT doing it on purpose*”). The Reviewer also proposes another fascinating idea – namely - that responding on the AMP may occur “even though I tried to stop it”. This would presumably involve a third factor (namely an active effort to combat the prime’s influence on one’s target evaluations; something that may require both awareness and intention to achieve).

Again, questions about the interaction between awareness, intention, and control over responding are interesting but beyond the remit of our paper. Nevertheless, we recognize the value of these ideas and have added new material to the General Discussion on this topic (see changes on p.66 in **blue**).

“*Exploring The (Interactive) Role of Automaticity Conditions Within the AMP. Our work focused on one particular automaticity condition (awareness), and one particular sub-type of awareness (influence awareness). It would be interesting to explore if other types of awareness also impact AMP effects in the way that influence awareness does. For instance, are AMP effects are moderated by a participant’s awareness of the content under investigation (i.e., that they have an automatic evaluative preference for one group over another). Or their awareness of how their automatic content is being assessed by the task (e.g., via speeded categorization of one social group with positive words and a second group with negative words as in the IAT). Or the presence of particular stimuli and response requirements in the procedure (e.g., that a black person is being presented as a prime stimulus). Or their awareness of the origin of their AMP effects (e.g., stemming from society or their personal learning history).*

*Similarly, it would interesting to explore if and how awareness interacts with other automaticity conditions, and how their interaction contributes to the AMP effect. Take, for instance, intention. Is it that awareness of the prime’s influence on target evaluations is a precondition for intentional responding (i.e., one can only intend to respond to the target in-line with the prime once they are aware of both stimuli and the latter’s relationship to the former)? Is it a pre-condition for controlling how one responds (demand) or for combatting the unwanted influence of certain stimuli over automatic behavior (e.g., if one wants to exert control over their automatic responses do they first need to be aware of the relationship between stimuli in the environment and how they influence those responses)? Is it possible to be influence aware and still respond in an unintentional manner? Future work could explore these and related issues.*”

**Reviewer 1**: I’d also like to see a bit more discussion about what is meant by “unaware”. Looking back to Bargh’s older work on this we might think of a participant being unaware of the prime, unaware of the direction of their evaluation of the prime, unaware of the strength of their evaluation of the prime, or unaware of the influence of the evaluation of the prime on the target. As I’m sure the authors know, more recent work from Hahn and Goedderz also raises questions of what it means to be “unaware”. I think it’s worth bringing this into the discussion for a more nuanced conclusion that “AMP effects are not implicit in the sense of being unaware”.

**Authors**: see our previous comment on this topic and footnote 15 which we have just added to the manuscript on p.66. We have also revised the conclusion section to better clarify our claim (i.e., “that AMP effects are not implicit in the sense of being (influence) unaware.”)

“*The different ways that awareness can be potentially defined and studied makes it difficult to (a) arrive at general conclusions about the role of awareness in the AMP from past work, and (b) requires precision about what type of awareness one is measuring, and how they have done so (for a more detailed treatment on the conceptual degrees of freedom surrounding awareness see Hahn & Goedderz, 2020; Corneille & Hütter, 2020). It is for this reason that we continually made reference to influence awareness throughout this paper, and sought to measure it using a range of measures (e.g., trial-by-trial vs. post-hoc self-reports), attitude domains, and both prospectively and retrospectively*)*.*”  
  
**Reviewer 1**: I’d really like to see simple correlations and descriptive statistics in the manuscript itself.

**Authors**: We have added simple descriptive stats in the manuscript as requested (see Tables 1 and 2). Also see the new addition of Figures 5-6.

**Reviewer 1**: In the latter studies, the authors do not report results from the “unaware” trials only – why not? This seems critical in that an AMP effect on only the unaware trials would suggest that the measure might be capturing \*some\* component of an evaluation that is outside of awareness? I think this info is in at least one graph but I think it could be laid out more clearly in the text itself.

**Authors**: In line with the Reviewer’s suggestion we have now include Means and SDs for AMP effects calculated on the basis of the “non-influenced” trials (see Table 2).

Briefly, we did find (a) very small IA-AMP effects when computed on the basis of only “non-influence” aware trials, and (b) these effects still possessed some predictive validity for discriminating between known groups. As the Reviewer suggests such a measure “may be capturing some component of an evaluation that is outside of awareness”.

However, as we note in the General Discussion, one should exercise caution making such a conclusion on the basis of *our* data. For instance, we would caution against conflating ‘non-influence aware’ with ‘influence unaware’ responding. The IA-AMPs we developed asked participants to press the spacebar if their evaluation was influenced by the prime. The presence of such a response provides a measure of influence awareness. Yet the absence of such a response is far more ambiguous.

It may be that such trials are free from influence awareness (i.e., are ‘influence-unaware’), or they could equally reflect uncertainty about influence, momentary distraction, or other sources of control over responding. Put simply, caution should be exercised when assigning a specific meaning to non-influence aware trials in the IA-AMP. To better investigate influence-unaware trials, one would need to develop and test a hypothetical ‘Influence-Unawareness AMP’ (IU-AMP): for example, by asking people to respond when their evaluation was not influenced by the prime.

We unpack this argument in even greater detail in the General Discussion.

**Reviewer 2**: This MS presents eight pre-registered studies testing whether participants are aware of being influenced by primes in the AMP, and, if so, whether this leads to differences in outcomes (AMP scores and relations with other variables). I found the main findings to be very convincing. The AMP is billed as an implicit measure, which, taken at face value, implies that participants should not be aware of the influence of the primes. If they are, and if - as the authors show - AMP effects are driven by trials where participants say they are influence-aware, that undermines the dominant interpretation of this measure (this interpretation is the one that the developers of the AMP have argued is correct, as the authors set out clearly in the introduction).

Of course, with sufficient creativity one can construct an alternative interpretation of these results on which (for example) participants are unaware of the influence of the primes when they respond but then infer their attitudes post-hoc. I don’t think any one alternative explanation covers all the data the authors report here. But, taking a step back, shouldn’t we have a low prior on such accounts just on the basis of parsimony? By far the simplest explanation for the current results is that people accurately reported when they were aware of the influence of the primes. As I understand it, the primary evidence for non-awareness is the Payne et al. (2013) skip-AMP vs. regular AMP null effect, which the authors very convincingly do not replicate in Study 1. So, I think the onus is on someone positing a more complicated account to give evidence for it.

**Authors**: We completely agree with Reviewer 2 in his assessment here.

**Reviewer 2**: My only real concern is about the choice of mixed-effects logistic regression to analyze most of the studies. I realize that the authors preregistered these analyses (which is great!) and therefore probably feel stuck with them. But I am not sure that the author’s research questions really require mixed effects; they are only interested in fixed effects as far as I can see. Why not use the simpler approach previously used to score the AMP, where you aggregate across trial type within person? (In this case you’d have to aggregate separately for influence aware and non influence aware trials, of course.) The upside of this is more interpretable results (I find differences in proportions easier to interpret than odds ratios) and less risk of misleading results because of model misspecification (which is a real risk with the more complicated mixed effects model, see cite below). If the authors do want to retain the mixed-effects models they ought to justify why they are not including random intercepts for primes and targets (by looking at ICCs or whatever the equivalent is for logit). It seems to me that you want to have random intercepts for any cluster variables as a default.

**Authors**: We appreciate and agree with the reviewer’s comments here. We wish to stick with our mixed-effects models as they are relevant to our analyses distinguishing subsets of trials vs. subsets of participants. However, we have now updated all of our logistic mixed-effects models’ random effects in line with the comments of the reviewer (see revised manuscript). Specifically, we now model a random intercept for prime identity in all models. As well as this, whereas previously we only modelled the random intercept of participant, we now also include random slopes for the prime type (i.e., to account for the fact that different participants may have different magnitudes of AMP effects). We believe our random effects structure is now in line with best practices for modelling implicit measures data.

**Reviewer 2**: Finally, I would love to see some simpler descriptive statistics, like what % of trials people reported influence awareness, what the range was across people, etc. This could probably go in one big table somewhere.

**Authors**: In line with Reviewer 2s suggestion we have included simple descriptive statistics in Tables 1 and 2. We have also included a third table outlining the number of trials people reported as being influence and non-influence aware on as a function of their AMP response (i.e., if it was a prime congruent or incongruent response).

**Reviewer 2**: Overall, despite these concerns I’m very positive about the paper. It is extremely thorough and I appreciated the preregistration and open data/materials, and the AMP is widely used enough that this will be of interest to many researchers.

**Author**: We sincerely thank the Reviewer for his kind words and constructive feedback.