Journal of Personality and Social Psychology: Attitudes and Social Cognition

Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness

--Manuscript Draft--

Manuscript Number:	
Full Title:	Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness
Abstract:	The Affect Misattribution Procedure (AMP) is used in many areas of psychological science based on the assumption that it not only taps into attitudes and biases but does so without a person's awareness. Across eight preregistered studies (N = 1603) plus meta-analyses we reexamined the 'implicitness' of AMP effects, and in particular, the idea that people are unaware of the prime's influence on their evaluations. Results indicated that AMP effects and their predictive validity are primarily moderated by a subset of influence aware trials (within individuals), and high rates of influence awareness (between individuals). Interestingly, an individual's influence awareness rate on one AMP predicted how they performed on an earlier AMP, even when the two assessed different attitude domains. Taken together, our results suggest that AMP effects are not implicit in the way that has been claimed, a finding that has implications for the procedure, past findings, and theory. All materials and data are available at osf.io/gv7cm.
Article Type:	Article
Keywords:	Affect Misattribution Procedure; Automaticity; implicit social cognition; implicit measures
Corresponding Author:	Sean Hughes, Ph.D Ghent University: Universiteit Gent Gent, Flanders BELGIUM
Corresponding Author E-Mail:	sean.hughes@ugent.be
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ghent University: Universiteit Gent
Other Authors:	Jamie Cummins
	Ian Hussey
Corresponding Author's Secondary Institution:	
First Author:	Sean Hughes, Ph.D
Order of Authors Secondary Information:	
Manuscript Region of Origin:	BELGIUM
Suggested Reviewers:	Yoav Bar-Anan Tel Aviv University baranan@tauex.tau.ac.il Prof. Bar-Anan is an expert in the field of implicit attitudes/measures and functioned as a Reviewer in our original submission.
	Christian Unkelbach University of Cologne: Universitat zu Koln christian.unkelbach@uni-koeln.de Prof. Unkelbach is an expert on implicit attitudes and measures.
Opposed Reviewers:	
Order of Authors:	Sean Hughes, Ph.D
	Jamie Cummins
	Ian Hussey

Re: PSP-A-2019-0728

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

Journal of Personality and Social Psychology: Attitudes and Social Cognition

Editor's comments

Dear Mr. Cummins,

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

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

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

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

Authors: We thank the Editor and Reviewers 1-2 for highlighting this issue to us. We apologize for any mischaracterization – it was not our intent. Many of these issues are no longer relevant as the content that was flagged is no longer present in this substantially revised version of the manuscript. That said, we still believe that one study on the unawareness of the AMP effect does contain a number of conceptual, statistical, and methodological concerns. We have done our best to accurate summarize and represent those concerns in a fair and balanced manner (for more see the revised introduction).

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

Authors: In line with the Editor and Reviewers 1-2's suggestions we have amended any reference to the idea that prime influence awareness "drives" AMP effects in the revised manuscript. We now instead state that the AMP effect is "strongly moderated by" awareness of influence of the primes, which removes the causal element of our language while still conveying a conditional relationship between AMP effects and awareness.

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

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

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

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

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

Authors: We have revised our paper to speak to the Editor's points in the following ways:

1. Our revised introduction unpacks the issues with past work on the implicitness of the AMP effect, avoids misunderstandings or misrepresentations, and clarifies the ways in which we addressed and improved upon what came before us.

2. We included a new, high-powered, replication attempt of a key study in support of the unawareness of AMP effects (Payne et al. 2013; Experiment 3). Results indicated that the original study (and thus claims) failed to replicate, further reinforcing the need for the role of awareness of AMP effects to be re-examined.

3. Two new empirical studies were carried out that extended our original analyses from retrospective to prospective awareness, and thereby providing a strong rebuttal of the post-hoc confabulation idea forwarded by Reviewer 2.

4. Analyses illustrating that a small subset of participants are responsible for the majority of variance in group-level AMP effects, and that these participants tend to be the same subset of people who are responsible for effects across different AMPs and across attitude domains. This represents another critical finding relating to the AMP: that AMP effects are not reflective of the population *in general* but rather a specific subset of people (the highly influence aware). To our knowledge this is the first finding of its kind, and represents an important contribution to the literature in terms of moderators of the AMP effect (and perhaps implicit measure effects more generally; for a detailed discussion see the General Discussion).

Reviewer 1's comments

Reviewer #1: Signed: Yoav Bar-Anan

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Finally, Reviewer 1 asked whether awareness might not occur at all unless directly prompted (as was the case in the IA-AMP). We directly speak to this point in Experiments 2-8. That is, if awareness was simply an artifact of our IA-AMP and unrelated to standard AMP effects then it should not have backward predicted standard AMP effects where awareness was never probed. Yet this is what occurred in all seven of our studies which directly tested this relationship (including one study where the same 'skip-AMP' paradigm was used as in Payne et al., 2013). Moreover, Experiment 5 also demonstrated that the relationship between influence awareness and AMP scores is *bidirectional*, insofar as AMP effect sizes predict how influence aware one will later report being, and how influence aware one is at an earlier point in time will predict the later magnitude of their AMP effects.

Thus we think it unlikely that awareness was simply a post-hoc irrelevance. If anything it appears to be a core factor predicting the size and predictive validity of AMP effects.

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

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

see the “Do AMP effects reflect a misattribution process?” section in our General Discussion).

On the other hand, however, we want to be clear. All of these possibilities are post-hoc justifications for the findings we report in our paper and not *a priori* claims made in previous empirical or theoretical papers. As far as we are aware, past work on the AMP effect and misattribution as a mental process do not claim that misattribution occurs *prospectively* and with *awareness* (see Experiments 7-8). Therefore, while we are happy to entertain post-hoc amendments to a theoretical concept, those post-hoc justifications should not be treated as equivalent to *a priori* pre-registered claims that conflict with them.

Instead, it should be acknowledged that an existing theoretical concept does not – as it stands – predict nor explain the pattern of findings we obtained across eight pre-registered, high-powered studies with different attitude domain (political vs. generic), variants of the AMP (standard, Mann et al., IA-AMP), and awareness measures (retrospective and prospective)

Rather what is being offered here is a post-hoc adjustment to a concept in order to modify that concept so that it can be retained in some form. But any such adjustment (i.e., that misattribution occurs prospectively with awareness) deviates significantly from what went before and should be empirically investigated rather than solely conjectured.

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

Authors: As noted above, Experiments 7-8 were designed to address this alternative interpretation of our original findings, and our results repeatedly contradict such an account. That said, we acknowledge this possibility in the discussions of those studies to provide a reminder to readers that alternative (post-hoc) perspectives are possible.

Reviewer 1: Related to the previous point, in p. 15, the authors wrote that they sought to determine if awareness drives AMP effects. They then use the verb "driven" often throughout the manuscript. I think that "drive" implies a causal role for awareness. However, the authors did not manipulate awareness. Therefore, they can conclude only about the possibility of a relation between awareness and the AMP effect, and not a causal relation. Very often, the word "drive" seemed inaccurate and might have conveyed the wrong message. Often, moderation of the priming effect by reported priming was described as evidence that the

priming effect was driven by awareness or by trials in which participants showed awareness, or by participants who reported much awareness. It is possible that I do not understand the meaning of "drive", but I do not think that it is common to describe findings of moderation, especially when the moderator is not manipulated, as evidence that the moderated effect is driven by the moderator.

Authors: As we note in our response to the Editor, we have toned down the causal language substantially in the manuscript and have removed mention of the word 'drive' throughout. We now state that AMP effects occur under the condition of awareness/that AMP effects are moderated by influence-awareness.

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

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

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

We are not claiming that the AMP effect does not meet any automaticity condition. For instance, target evaluations are typically emitted quickly in the task (thus satisfying the speed condition). To our knowledge the 'goal-directed' nature of the effect has not been investigated, so it may be the effect is implicit is that sense too. We simply argue that

‘implicit’ is an umbrella term of a set of automaticity conditions, and that the AMP effect meets some of these criteria (speed) but not others (awareness). In this way it is ‘implicit’ in some ways and not others.

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

Authors: We thank Reviewer 1 for pointing this omission out to us. We have revised the introduction to include mention of this aspect of Payne et al.’s (2013) Experiment 2 design.

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

Authors: See our previous response.

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

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

Authors: We respectfully disagree with the Reviewer and maintain that our original argument here holds. The original version of the skip-AMP provides incomplete data insofar as it requires participants to *either* evaluate the target (i.e., provide evaluative information) *or* indicate that their response would have been influenced by the prime (i.e., influence aware information). But never both. As such, it is impossible to directly compare performance on trials where people indicated that they were influenced to those trials where they reported no

such influence. Without both pieces of information, it is difficult to determine what impact influence-aware trials have on the AMP effect, and if this impact is comparable to, or greater than, that of the non-influenced trials.

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

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

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

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

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

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

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

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

Reviewer 1: It was not entirely clear what methodological shortcomings Gawronski & Ye's (2015) research had. Their crucial finding was that the retrospective reports of the priming effect correlated with the priming effect only for the topic that was salient during the task, and not for the topic that was not salient. If the reason for the correlation between the priming effect and retrospective reports of the priming is due to intentional rating of the primes, why would the manipulation of topic salience influence this correlation without influencing the priming effect itself? The present authors wrote "retrospective self-reports do not provide a direct assessment of the construct under investigation". Yet, Gawronski and Ye did not rely on those self-reports as a measure of awareness of the priming effect. Rather, they tested whether the finding of a correlation between retrospective self-report and the priming effect

survives a certain manipulation of awareness. They showed that their manipulation of awareness decreased the validity of the self-reported awareness of the priming effect as a predictor of the priming effect but did not decrease the priming effect itself (the results summarized in Table 1 in Gawronski & Ye's article are the best evidence I have seen so far, against the intentional rating account). It seems reasonable to conclude from that evidence that the self-reported awareness of the priming was not due to a necessity of awareness for the priming effect to occur.

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

Reviewer 1. The authors conclude that the AMP priming effect "do not represent an equally valid measure of attitudes across individuals". This seems a valid conclusion from the evidence they report, and it is compatible with the evidence reported in Bar-Anan & Nosek (2012, 2014). In our 2012 research (mainly in Tables 3 and 4), we showed that indices of psychometric quality are reduced when excluding from the analyses participants who reported intentional rating of the primes (or, at least, awareness of the priming effect). We also found (see Appendix D of Bar-Anan & Nosek, 2014, Figures A and B, at https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-013-0410-6/MediaObjects/13428_2013_410_MOESM1_ESM.pdf) that the AMP loses its relation with direct measures of evaluation much faster than other indirect measures, after removing participants with extreme scores (those with the largest priming effects). However, all that evidence is still insufficient to inform us how serious this problem is. Only the appendix from our 2014 paper provides some comparison with other indirect measures (and the AMP seems inferior to the other measures). Yet, I did not see much research about how many participants "drive" typical effects in social psychology, and how many are the main contributors to validity evidence of psychological measures. I also do not know of much research that informs us how inequality in validity of a measure across individuals affects scientific progress. Clearly, it is better if a measure works well for a larger portion of the population, but what is the standard and how much does scientific progress suffer from each drop in that equality? I think that readers would need that knowledge in order to make strong conclusions about the implications of the inequality reported in the present manuscript.

Authors: We are glad that Reviewer 1 agrees with us that our evidence highlights that the AMP effect is not an equally valid measure across all individuals. We also agree with the reviewer that at present it is unclear how the AMP measures up to other measures within psychology in terms of its heterogeneous validity across participants. We agree that this issue is worth exploring in the context of other (implicit) measures and explicitly address this in our revised General Discussion. Unfortunately, there is currently little-to-no data available now that would allow us to make such a comparison.

Reviewer 1: In the "Structural Validity" section, the authors seem to expect a negative correlation between rating of targets after positive primes and rating of targets after negative

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

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

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

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

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

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

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

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

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

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

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

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

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

Authors: Reviewer 1 is correct that the Supplementary Materials refer to the html Markdown files produced by the analysis files. Note that outputting Word or PDF versions of these files is possible through the use of RMarkdown within our original analysis files.

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

Authors: The paper has been revised as requested (see meta-analysis section).

Reviewer 1: In p. 45, to interpret the moderation of the priming effect in each trial, by the self-reported awareness of the priming effect, the authors compared the moderation effect-size

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Authors: We have revised the description of the figures throughout the manuscript to explicitly describe what it is the x-axis labels refer to.

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

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

Reviewer 2's comments

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

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

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

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

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

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

Authors: See our previous comment.

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

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

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

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

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

Reviewer 2: Another way to look at this problem is that all of the analyses depend on the correlation between reports of awareness and the priming effect. The authors interpret their findings as evidence that people who show systematic priming effects have disregarded the instructions and intentionally rated the targets consistent with the primes. That is, aware and intentional ratings cause the priming effects. But all of the findings are just what the misattribution account predicts also. The misattribution account says that it is difficult to disentangle affective response to the primes and targets, so subjects often mistake the source of the affect as the pictograph target when it is actually the prime. (A misattribution by

definition can't be made with awareness or intention). Participants can observe their own behavior and notice if they are responding in prime-consistent ways. If so, they can report afterward that they were influenced by the prime (see Payne et al, 2013 for the same argument). This means that when priming effects are larger, subjects should report more influence of primes. If you divide subjects into those that reported large influences and those who didn't, then those who did not report influence won't have much priming because they have been selected to be that way. So these studies do not distinguish between the misattribution account and the authors' intentional/aware account at all.

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

Post-hoc confabulation

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

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

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

Misattribution

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

If one wants to explain AMP effects in these two studies in terms of misattribution, then they would need to allow for the idea that people are not only *aware* of misattribution but also able

to *predict* that it is going to occur before a target is evaluated or a target stimulus is even presented onscreen. Yet such an approach runs contrary to how misattribution is traditionally defined (Schwarz & Clore, 1983), and would require a radical overhaul of the concept itself.

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

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

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

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

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

Reviewer 2: In the introduction the authors attempt to argue against some of the previous points made in the exchange between Bar-Anan and Nosek and Payne et al (2013) and Gawronski and Ye (2014, 2015). First, they argue that it is problematic that the AMP defines what is intentional and unintentional by the instructions, and they note that sometimes subjects don't follow instructions and instead incorporate information that the researchers

instruct them to ignore (p. 11). Subjects sometimes do this, of course, but the question at issue is why. Unintentional effects of primes on judgments is one reason they do so, although there are of course other reasons. Nonetheless, using instructions to define intentional responding is not a weakness. In fact, virtually every task that aims to measure performance by accuracy and errors must use instructions to define task goals and therefore what is accurate or error, and what is intended vs. unintended responding. For example in the Stroop task, experimenters must use instructions to tell subjects to name the font rather than read the words. Responses that diverge from the task goal (which is set by instructions) define automatic or unintentional behavior.

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

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

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

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

Authors: We recognize that our characterization of the Payne et al. study was incorrect on several fronts. We sincerely apologize for those errors and have revised the manuscript to correct this (and other) such issues (see our reply to Reviewer 1's comment and the revised manuscript).

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

Reviewer 2: Next the manuscript says "Critically, however, the inference that 'intentional' AMP effects were "more affected" (p. 381) by the race of the prime than 'unintentional' AMP effects was never directly addressed in any of their other analyses…" and then go one to say we should have tested an interaction rather than reporting that an effect on one version of the test was significant and the other was not. But the present authors are entirely mistaken about the analyses we reported, and so their criticism is uninterpretable. That study examined the

associations between two forms of the AMP (an indirect version in which subjects judged the pictograph targets and a direct one in which they were instructed to rate the primes) and impression judgments of a black or white target character (we examined main effects and interactions in a regression framework). And we tested the effect of seeing the black target character versus the white target character on indirect and direct AMP tasks. The hypothesis tested was that when people intentionally rate the primes their responses will be more reactive than the indirect version to the task they just completed. It is not clear how to respond to the statistical issues raised in this section given that the errors make it difficult to know what the authors are talking about.

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

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

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

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

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

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

Authors: We appreciate Reviewer 2’s comments. We now feel that the addition of our two new experiments addressing the question of post-hoc confabulation within our results, combined with our third new experiment attempting to replicate the findings of Experiment 3

of Payne et al., establish even more clearly the contribution which our work can make to the field and the question of awareness in the AMP.

Cover Letter

Sean Hughes

E Sean.Hughes@UGent.be
www.drseanhughes.com

Faculty of Psychology and Educational Sciences
Henri Dunantlaan 2
B-9000 Ghent
Belgium

DATE	PAGE	OUR REFERENCE
16/03/2021	1	Manuscript Submission

Dear Professor Berkman,

Please find attached our revised manuscript titled "*Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness*", which was invited for resubmission as an empirical paper in the "Attitudes and Social Cognition" section in *Journal of Personality and Social Psychology* (PSP-A-2019-0728).

First and foremost, we would like to thank you and Reviewers 1-2 for your constructive feedback. We took that feedback seriously and used it to substantially revise our manuscript. It led us to carry out three new high powered, pre-registered studies that speak directly to the issues raised during the initial round of reviews. We believe these empirical additions, as well as the extensive revisions to our manuscript, address the original concerns and have resulted in a far stronger manuscript. In addition to these large-scale changes we have carefully considered each of your comments as well as those of the two reviewers. You can find an overview of our responses in an attached document.

All authors have approved the current version of the manuscript and made significant contributions to its conceptualization, statistical analyses, and/or writing. The manuscript meets the guidelines for ethical conduct and reporting of research, and holds no potential or actual conflicts of interest. It is not under review elsewhere; the data have not been previously published or accepted for publication.

Kind Regards,

Sean Hughes (sean.hughes@ugent.be)

Corresponding author

Jamie Cummins (jamie.cummins@ugent.be)

Ian Hussey (ian.hussey@ugent.be)

Co-authors

Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness

Sean Hughes, Jamie Cummins, & Ian Hussey

Ghent University, Belgium

Author note

All authors contributed equally to the manuscript and agree to be joint first author. SH, JC, and IH, Department of Experimental Clinical and Health Psychology, Ghent University. This research was conducted with the support of Grant BOF16/MET_V/002 to Jan De Houwer and Ghent University postdoctoral fellowship 01P05517 to IH. Correspondence concerning this article should be sent to sean.hughes@ugent.be, jamie.cummins@ugent.be, or ian.hussey@ugent.be.

Abstract

The Affect Misattribution Procedure (AMP) is used in many areas of psychological science based on the assumption that it not only taps into attitudes and biases but does so without a person's awareness. Across eight preregistered studies (N = 1603) plus meta-analyses we reexamined the 'implicitness' of AMP effects, and in particular, the idea that people are unaware of the prime's influence on their evaluations. Results indicated that AMP effects and their predictive validity are primarily moderated by a subset of influence aware trials (within individuals), and high rates of influence awareness (between individuals). Interestingly, an individual's influence awareness rate on one AMP predicted how they performed on an earlier AMP, even when the two assessed different attitude domains. Taken together, our results suggest that AMP effects are not implicit in the way that has been claimed, a finding that has implications for the procedure, past findings, and theory. All materials and data are available at osf.io/gv7cm

Keywords: Affect Misattribution Procedure; automaticity; implicit social cognition; implicit measures

Effects on the Affect Misattribution Procedure are Strongly Moderated by Awareness

Over the past twenty years research on implicit cognition has grown from a relatively niche field into, what is today, one of the most prolific and widely examined topics in psychological science. The idea that our automatic thoughts, feelings, and actions shape downstream behavior drives research, theory, and application throughout the discipline, especially in social and personality psychology, neuroscience, health, cognitive, and clinical psychology (for a book length treatment see Gawronski & Payne, 2010).

The success of the topic has been due in large part to the development and widespread use of tasks known as *indirect measurement procedures*. In contrast to *direct measurement procedures*, which simply ask people to report on their thoughts, feelings, and actions, indirect procedures seek to probe the mind by interpreting performance (e.g., speed and/or accuracy) on experimental paradigms. Notable examples include the Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998), evaluative priming tasks (Hermans, De Houwer, & Eelen, 1994), and approach-avoidance tasks (Rinck & Becker, 2007; for a review see Gawronski & De Houwer, 2014). The outcomes of these procedures are commonly referred to as *implicit measures* (e.g., the IAT *effect*, priming *effects*; for more see De Houwer, 2006).

Indirect procedures are often deployed under the assumption that they limit a person's ability to control how they respond, or their need for introspective access and/or conscious awareness of the content under investigation (i.e., that they operate under the conditions of automaticity). As a result, these tasks are typically used when researchers want to gain insight into content that people may be unwilling or unable to report (see Greenwald et al., 1998; Hahn & Gawronski, 2019). Although debate continues about what implicit measures actually reflect (Brownstein, Madva, & Gawronski, 2019; Corneille & Hutter, 2020; Schimmack, 2021), a vast

and ever-increasing number of studies continue to rely on indirect procedures and their effects to provide insights that other (self-report) procedures cannot.

The Affect Misattribution Procedure

The Affect Misattribution Procedure (AMP) has emerged as one of the more popular indirect procedures because it possesses the structural advantages of sequential priming along with good psychometric properties that other indirect procedures often lack (see Payne & Lundberg, 2014). At its core, the AMP consists of trials made up of three elements: (a) a prime stimulus (e.g., an image of a social in-group member) which is first flashed on screen for a brief period of time, followed quickly by (b) a target stimulus (usually a neutral Chinese pictograph), which is subsequently masked by (c) a white noise image. The AMP requires participants to subjectively evaluate how visually pleasing the target stimulus is, while ignoring the prime that preceded it. Despite being explicitly told to disregard the prime when evaluating the target, people nonetheless evaluate the latter in ways that are consistent with the valence of the former. For instance, when a neutral Chinese pictograph is preceded by a social in-group member, people are more likely to evaluate it as pleasant, compared to when it is preceded by a social out-group member (Payne, Cheng, Govorun, & Stewart, 2005).

Since its creation, the AMP has attracted considerable attention. It is commonly used in social psychology to assess attitudes in domains such as race (Payne et al., 2005; Ditonto, Lau, & Sears, 2013; although see Teige-Mocigemba, Becker, Sherman, Reichardt, & Klauer, 2017), gender (Ye & Gawronski, 2018), sexuality (Imhoff, Schmidt, Bernhardt, Dierksmeier, & Banse, 2011), and politics (Payne et al., 2005; Kalmoe & Piston, 2013). It has been used to investigate the origins of attitudes and stereotypes (Dunham & Emory, 2014; Mann et al., 2019; Van Dessel, Mertens, Smith, & De Houwer, 2017), and to assess the effectiveness of attitude change

interventions (Mann & Ferguson, 2017). In clinical psychology, it is used to assess, and sometimes provide prospective prediction of, maladaptive behaviors such as eating disorders, non-suicidal self-injury, alcoholism, anxiety, depressive symptoms, and physical abuse (Fox et al., 2018; Görgen, Joormann, Hiller, & Witthöft, 2015; Jasper & Witthöft, 2013; McCarthy, Skowronski, Crouch, & Milner, 2017; Smith, Forrest, Velkoff, Ribeiro, & Franklin, 2018; Zerhouni, Bègue, Comiran, & Wiers, 2018). Some clinical researchers also use the task as an outcome measure to benchmark the effectiveness of psychological interventions (Chapman et al., 2018; Schreiber, Witthöft, Neng, & Weck, 2016).

Two Competing Accounts of the AMP Effect

Two distinct perspectives have emerged to explain the aforementioned effects: an *implicit* account and an *explicit* account.[1] Both start from the position that AMP effects represent a valid measure of attitudes and bias. However, they differ in how "implicit" or "automatic" those effects are said to be. On the one hand, the implicit account argues that AMP effects reflect evaluations captured under certain conditions of automaticity (i.e., specifically, in the absence of both intention and awareness; Payne et al., 2005; Payne et al., 2013). On the other hand, the explicit account rejects this idea and argues that participants are aware of the prime's influence on their evaluations, and exert intentional control over their behavior in order to respond in-line with those primes (Bar-Anan & Nosek, 2012; Mann et al., 2019). In what follows we briefly consider research which has examined the issues of awareness and intention of AMP effects.

[1] 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 (see Moors & De Houwer, 2006). 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. For those looking for an extensive debate about the meaning and usefulness of the term implicit we recommend Corneille and Hütter (2020).

Intentionality in the AMP

Evidence for the explicit account mainly comes from Bar-Anan and Nosek (2012) who asked participants to first complete an AMP and afterwards indicate if they had intentionally based their evaluations on the prime rather the target. 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).

Proponents of the implicit account conducted several experiments which rejected these claims. For instance, Payne and colleagues (2013) found that the relationship between intentionality ratings and AMP effects was similar when people had to indicate if they were intentionally or unintentionally influenced by the prime. Drawing on this finding they claimed that 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). In a second experiment, participants were asked to complete the AMP twice: once where they had to evaluate the target instead of prime (standard 'unintentional' AMP) and once where they had to evaluate the prime instead of the target (an 'intentional' AMP)[2]. The authors found that the relationship between standard AMP and personality judgements of a Black person were different to the relationship between the intentional AMP and that same personality judgement. These results and others (e.g., Gawronski & Ye, 2014) have been advanced in support of the idea that AMP effects are unintentional in nature. Notably, recent work has been advanced in favor of the intentional nature of AMP effects (Bar-Anan & Nosek, 2016; Mann et al., 2019). In short, there have been a number of studies investigating intentionality within the AMP, and there is currently no universal consensus on whether AMP effects qualify as unintentional.

[2] Payne et al. (2013) referred to these as these 'indirect' and 'direct' AMPs, respectively.

Awareness in the AMP

The implicit and explicit accounts also 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. However, unlike intentionality, awareness within the AMP has been subject to comparably little empirical investigation.

One study to address this issue was conducted by Payne and colleagues (2013; Experiment 3). They divided participants into two groups: the first completed a standard AMP, whereas the second completed a 'skip' AMP. During the latter AMP participants were given the option to respond in one of three ways: they could either indicate that the target stimulus was pleasant, unpleasant, or choose to 'skip' that trial entirely if they felt that their evaluation would have been influenced by the prime. The authors argued that if AMP effects were due to responding on trials where participants were aware of the prime's influence on their evaluations, then removing such trials "should eliminate the priming effect" (p. 377). When they compared skip-AMP effects (where influence aware trials had been removed) to standard AMP effects they found that the former did not significantly differ from the latter.

Awareness Revisited

In light of Payne et al. (2013, Experiment 3), it may be tempting to conclude that AMP effects occur without awareness and that are therefore implicit in this manner. We disagree. Such claims may be premature given that the aforementioned study is, in our opinion, subject to a number of issues which impact the interpretations originally made, which we will now discuss.

Methodological Issues

On the surface the 'skip' AMP developed by Payne et al. (2013) appears to provide an *in vivo* measure of awareness insofar as participants are provided with an option to signal that the prime has influenced their evaluations. However, this task has its issues. Perhaps, most importantly, it requires people to make an either/or decision: *either* provide an evaluative response *or* indicate that they were aware of the prime's influence. But it never does both (i.e., allow the participant to respond to the target *and* indicate this was a 'contaminated' response). As such, it is impossible to directly compare performance on trials where people indicated that they were influence aware to those trials where they were non-influence aware. Without both pieces of information, it is difficult to determine whether trial-by-trial influence awareness moderates the magnitude of the AMP effect, or how AMP effects calculated from only influence aware trials compare to those calculated from non-influence aware trials.

Statistical Issues

Payne et al. (2013; Experiment 3) argued that effects on the standard AMP did not differ from those on the 'skip' AMP, and used this as evidence in support of the implicit account. Yet this conclusion is also questionable given that non-significant statistical differences between two means does not necessarily imply that they are statistically equivalent (Lakens, Scheel, & Isager, 2018; Quertemont, 2011). As such the original inference drawn was not supported by the analyses conducted.

Conceptual Issues

In Experiment 3 of Payne et al. (2013), the authors noted that participants tended to skip trials more frequently on trials with neutral compared to valenced primes. They suggested that such a pattern could be explained by the implicit but not by the explicit account (i.e., that if

people were aware they should skip when confronted with valenced primes and not with neutral primes). We disagree. The explicit account 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 determine their response to the target. In cases where the prime is neutral there is no evaluative information available which one can use to guide their response to the target. Thus it follows that they will skip more on such trials. The opposite is true on valenced prime trials and thus skipping should occur less frequently.

Conclusion

Only one study to date has investigated the question of awareness in the AMP. Given the conflicting accounts of the role of unawareness within the AMP, the comparably little empirical attention it has received, and the combination of methodological (absence of information about influence aware trial performance), statistical (conflation of statistical non-significance and statistical equivalence), and conceptual issues (equally plausible explanation of findings by the explicit account) within this study, it seems reasonable to examine in greater depth whether participants really are aware of the prime's influence on their evaluations.

The Current Research

Across eight preregistered studies (1 replication and 7 novel studies) we re-examined the implicitness of AMP effects and, in particular, the assertion that people are unaware of the prime's influence on their evaluations. In Experiment 1 we conducted a high-powered, preregistered replication of Payne et al.'s (2013, Experiment 3) work with the 'skip' AMP. The finding that 'skip' AMP effects are no different to standard AMP effects is viewed as strong support for the implicit account. To briefly preface what is to come, the authors' original

findings did not replicate, such that scores on the standard AMP were significantly larger than those on the skip AMP, undermining the implicit account.

In Experiment 2 we sought to address a key limitation of the original 'skip' AMP - namely - that it forces people to either skip *or* evaluate the target and thus only provides partial data. We developed an influence aware (IA-) AMP that had participants rate the target (provide evaluative information) and then indicate if evaluations had been influenced by the prime (provide influence information). We found that AMP effects were moderated at the trial-by-trial level by influence awareness, as well as by at the group level by inter-individual differences in influence awareness.

In Experiments 3 and 4 we controlled for the possibility that by probing for influence awareness on each trial of the IA-AMP we artificially altered the relationship between awareness and AMP effects. Participants now completed a standard AMP at Time 1 and an IA-AMP at Time 2, either from the same (Experiment 3, i.e., both generic valence) or different attitude domains (Experiment 4, i.e., one generic valence and one politics). In both cases influence awareness during an IA-AMP at Time 2 predicted the magnitude of standard AMP effects at Time 1, indicating that influence awareness is a stable (within-participant) pattern of responding that holds within and between content domains.

In Experiment 5 we had two groups of participants (Democrats and Republicans) first complete a political IA-AMP and then an IA-AMP with generic valenced primes. We found that the AMP's ability to correctly classify a person as a Democrat or Republican was superior when effects were based solely on influence aware trials and inferior when based solely on non-influence aware trials. Experiment 6 had participants first complete a newly developed version of the AMP that purportedly reduces subset effects within the AMP (the Mann et al. [2019]

modifications to the AMP) followed by a Mann et al. IA-AMP. Once again, the same pattern of findings emerged as outlined above, even within a variant of the task designed to optimize the implicitness of the AMP.

In our final two studies we modified the IA-AMP so that influence awareness was measured *prospectively*, either before the target was evaluated (Experiment 7) or before the target stimulus was even presented (Experiment 8). In this way influence awareness was measured before an overt evaluation took place or a covert evaluation could even be formed. 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 no second stimulus to misattribute evaluations to or confabulate awareness from at the point within the task when awareness of influences was reported (see Figure 1).

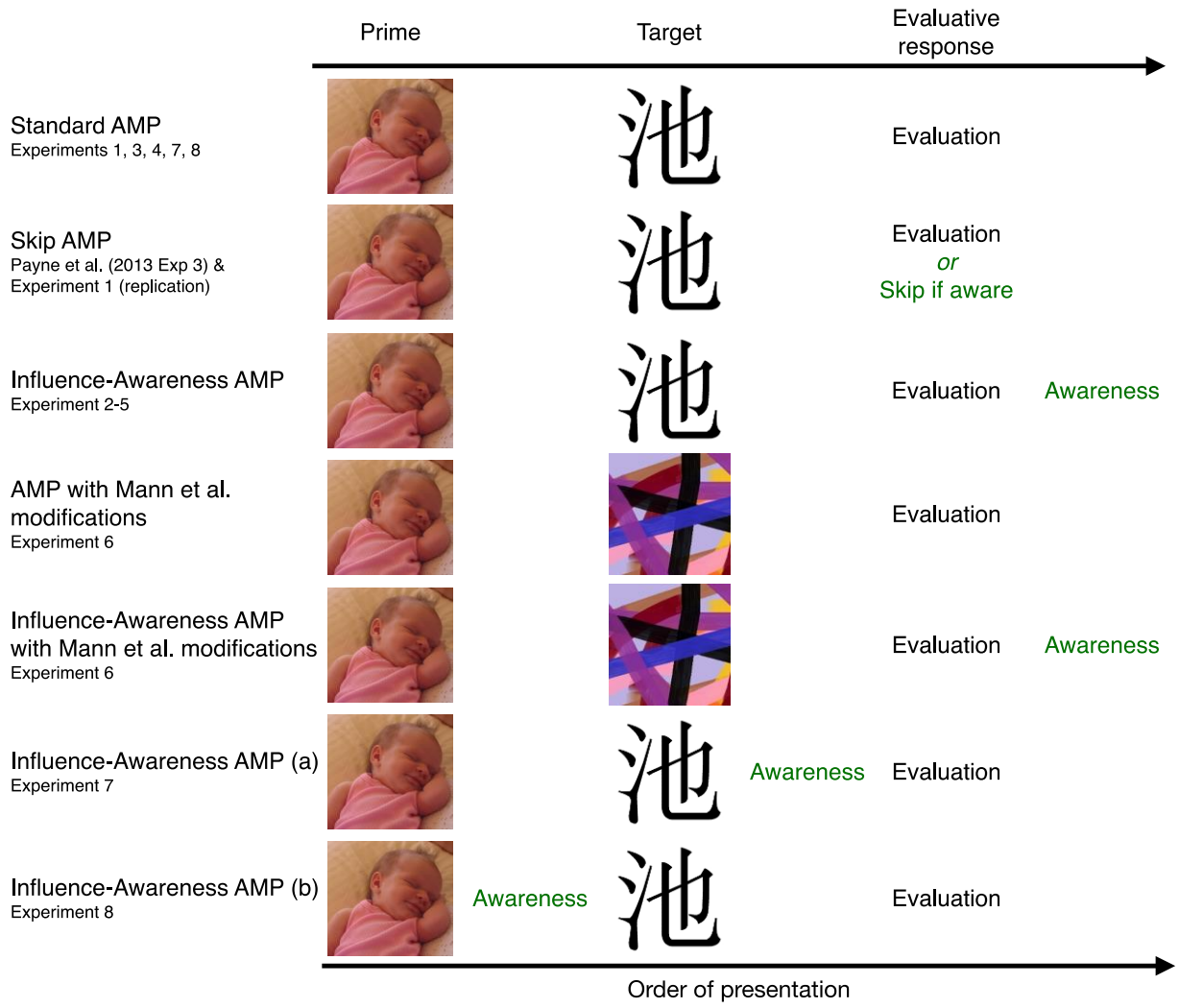

Figure 1. A schematic overview of the AMP variants used in Experiments 1-8. The point within the trial at which awareness of influence of the prime on evaluations was systematically varied between studies.

Experiment 1: Payne et al. (2013, Experiment 3) Fails to Replicate

In their original paper, Payne et al. reasoned that "if a participant is aware when she is being influenced by a prime, then she can pass when she would otherwise display a priming effect. The trials on which he or she chooses to forego the pass option and evaluate the

pictograph should therefore be free of influence from the primes. If subjective experiences of being influenced by the primes are well calibrated to actual influence, then the pass option should allow respondents to eliminate the priming effect." (p.382). In other words, if awareness of influence of the prime is central to AMP effects then effects on the standard version of the task should be larger than effects on the skip variant. The authors found no statistically significant difference between the two task variants and concluded that "these data contradict the idea that participants were aware that the primes influenced them before responding" (p.383).

In Experiment 1 we examined if this claim (that 'skip' AMP effects do not differ from standard AMP effects) is replicable. We first carried out power analyses (detailed below) to ensure that we had sufficient power to detect even small effects (something that may have presented a problem in the original study).[3] We then administered similar standard and skip-AMPs as used by Payne et al. (2013). Whereas the original authors relied on a between- subjects manipulation we opted to administer both AMP variants to all participants in order to improve our statistical power, as well as to better compare effect sizes within rather than between individuals. We also carried out statistical comparisons that allowed us to test the original key claim (e.g., we used a partially-overlapping *t*-test to account for participants who skip either all trials or no trials). In addition to the confirmatory analysis, we also asked an exploratory question: does one's awareness of the prime's influence on their evaluations (as indexed by skip rates in the skip AMP) predict the magnitude of their effect in a standard AMP? If so, then this

[3] The sample size used in the authors original study was relatively small (N = 36 per cell). Although they argued that this sample would provide power "greater than .85 to detect an interaction between the repeated measures and between-subjects factors, even assuming a small effect size" (p. 383), they did not specify what effect size they qualified as "small", nor what they specified as the correlation between the different within-subject measurements (i.e., the correlation between evaluations of the positive and negative primes). This unspecified correlation can make such a power analysis vary widely. Combining this with the fact that their sample size was relatively small, it is very possible that a true difference between the AMP types exists, but the authors simply did not achieve sufficient power to detect such a difference.

would suggest that influence awareness may play more of a role in standard AMP effects than previously thought.

Method

Materials for all experiments can be found at osf.io/gv7cm. This includes details of the designs, experimental scripts, raw and processed data, preregistrations, analytic plans, and all R code for data processing and analyses. We also report how sample sizes, data exclusions (if any), manipulations, and measures were determined in each study. We report only the key effects that serve to test our hypotheses. All other results of the models can be found on OSF.

Sample Selection Strategy

Power analyses indicated that 147 participants would be required to detect a Cohen's *d* effect size of 0.3 in a paired-samples *t*-test at the conventional alpha level (.05, two-sided) with 95% power. 289 participants would be required to detect such an effect size in a two-sample *t*-test with otherwise identical parameters. Given that a partially-overlapping *t*-test's power typically falls somewhere between a paired-samples *t*-test and a two-sample *t*-test (Derrick, Toher & White, 2017), 289 participants would provide (a) at least 95% power to detect a small effect size in such a test and (b) power to detect a very small Cohen's *f* effect size (i.e., 0.045) in a linear regression with one dependent variable and one independent variable (i.e., the analysis used to investigate our second question).

Participants and Design

316 individuals were recruited via Prolific (prolific.co) and took part in exchange for a monetary reward. We initially recruited 290 participants, but a number provided incomplete or partial data or did not met our preregistered exclusion criteria. Recruitment was continued in batches of 10 until analyzable data was available for at least 290 individuals (final n = 295; 160

men), who ranged in age from 18 to 61 (M = 29.8, SD = 10.3). A 2 (*Task Type*: standard vs. 'skip' AMP) x 2 (*Prime Type*: positive vs. negative) design was employed with both factors manipulated within participants. Ratings of the target stimuli (positive and negative images) served as the dependent variable.

Ethical Approval

Approval for all studies was provided by the Ethical Committee of the Faculty of Psychology and Educational Sciences at Ghent University (approval numbers 2015/13, 2016/63, and 2016/80).

Materials

Materials were programmed in Inquisit 4.0 and administered via the Inquisit Web Player. Both versions of the AMP contained three types of stimuli: primes, targets, and a mask. Prime stimuli consisted of 12 positive and 12 negative images taken from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1997). Target stimuli consisted of 72 Chinese pictographs and the mask consisted of a white noise image.

Procedure

Participants initially provided informed consent and demographic information (age and gender) and then completed the standard AMP followed by the skip-AMP.

Standard AMP. Prior to the task participants were told that pictures would appear one after another on the screen. The first would be a real-life image and the second a Chinese symbol. Their task was to judge the visual pleasantness of the Chinese symbol using the E (pleasant) and I (unpleasant) keys while trying their best to not let the real-life images bias those judgements. Overall, the task consisted of 72 trials. Each trial began with the presentation of a positive or negative prime stimulus for 75ms, followed 100ms later by a target image (i.e., a

Chinese pictograph) which remained onscreen for 100ms, after which, a white noise image appeared and replaced the Chinese pictograph. This mask remained onscreen until the target stimulus was evaluated as positive or negative using the E or I keys respectively.

Skip AMP. The skip AMP was similar to the standard AMP. Participants were informed that they would complete a similar task once again and that they could now respond in a third way - namely - to 'skip' a trial by pressing the spacebar whenever they thought their evaluations of the pictographs might be influenced by the prime. Instructions emphasized that they should only evaluate the pictograph whenever their opinion reflected the qualities of the pictograph itself. The AMPs used in our replication were similar to those used by Payne et al. (2013) with two exceptions. First we use 72 rather than 120 trials in order to make completion of two AMPs manageable for participants. Second, whereas Payne et al. used valenced and neutral primes we only used valenced primes as - in most AMP studies - only valenced primes are used.

Results

Data Exclusion

Participants were excluded if they completed the experiment too quickly (i.e., in under three minutes) or provided incomplete data on any of the measures (n = 21).

Data Preparation

AMP effects were computed by subtracting the proportion of 'pleasant' responses emitted on trials with an unpleasant prime from the proportion of 'pleasant' responses emitted on trials with a pleasant prime (Payne et al., 2005). Scores were calculated from all trials in the standard AMP and exclusively from the non-skipped trials in the skip-AMP.

Analytic Strategy

We carried out a partially-overlapping *t*-test to examine our first question (i.e., do AMP effects differ as a function of *Task Type* [standard vs. skip]). We opted for this test for the following reason: it may be that some participants recognized the prime's influence on their evaluations and therefore skipped all trials in the skip-AMP. If so, then these individuals produced no AMP scores on this version of the task. One could simply exclude such participants in order to run a paired-sampled *t*-test between skip and standard AMP scores. Yet skip-AMP effects are not missing at random and are instead missing for a very important reason (i.e., people are highly influenced). Excluding such individuals would undermine the inferences we ultimately want to make.

The partially overlapping *t*-test is a variant of the *t*-test which overcomes this issue (Derrick et al., 2017). It is neither a dependent nor independent *t*-test but rather a mixed *t*-test containing independent and dependent data. Given that participants with an influence awareness rate close to 100% had no skip AMP effect and therefore had no data for the skip AMP, their standard AMP effects were entered as independent data. Those with standard *and* skip AMP effects were entered as dependent data.

We also investigated a second preregistered exploratory question: whether influence awareness rates in the skip AMP predict the magnitude of effects in the standard AMP. To answer this question, we carried out a linear regression analysis with rate of skipping in the skip-AMP as the independent variable and effects in the standard AMP as the dependent variable. All reported analyses were preregistered in this and all other studies unless otherwise noted.

Hypothesis Testing

Do Effects On The Standard AMP Differ From Those On The Skip-AMP? Results indicated that effects in the standard AMP were significantly larger than those in the skip-AMP,

$t(294) = 9.36$, $p < .001$, $M_{diff} = 0.29$. A between-subjects Cohen's *d* was also calculated for familiarity, although this should be interpreted with caution as it does not acknowledge the partial dependance within the data: $d = 0.96$, 95% CI [0.79, 1.13].

Does The Rate of Skipping in The Skip AMP Predict Standard AMP Effects?

Regression analyses indicated that rates of skipping in the skip AMP predicted the magnitude of effects in the standard AMP, $B = 0.30$, 95% CI [0.07, 0.53], $\beta = .15$, 95% CI [0.04; 0.26], $p = .010$.

Discussion

A high-powered, preregistered replication attempt of Payne et al. (2013; Experiment 3) was unable to replicate their findings. Whereas the original authors found no difference between a skip- and standard AMP we found a difference between the two, such that scores on the standard AMP were significantly larger than those on the skip AMP. Contrary to the original authors' claims, it seems that people's subjective experiences during the skip-AMP were well-calibrated to the actual influence of the primes on their responses, and this allowed them to significantly reduce the priming effect. Even more interestingly, we found that a given participant's awareness of the prime's influence on their evaluations (during the skip-AMP at Time 2) strongly predicted the magnitude of their effects in the standard AMP at Time 1. [4] This suggests that awareness of the prime's influence on evaluations may play a role in the standard AMP as well.

[4] Throughout this article we employ the word 'predict' in its statistical sense, i.e., the estimate scores on one task based on scores on another task. Our claims are agnostic to the temporal order of the tasks, which are based exclusively on the statistical relationship among scores on the tasks. Indeed, the temporal order of the tasks in our experiments were specifically chosen so that participants complete a standard and unaltered AMP before other tasks so as to exclude the possibility that performance on the standard AMP was altered or perturbed in any way by the other tasks. Elsewhere, when the dependent variable is completed prior to the independent variable, this is occasionally referred to as 'postdiction' rather than 'prediction', however we employ the latter for familiarity.

Experiment 2: AMP Effects are Strongly Related to Awareness of the Prime's Influence on Evaluations

Although the skip-AMP is an improvement on post-hoc awareness measures it is not without its issues. By forcing people to either skip or evaluate the target stimulus the task can only provide partial data (i.e., either evaluations *or* indications of influence). A superior task would be one where participants provide their evaluation of the target *and* indicate if that evaluations was influenced by the prime. Such a task would provide evaluative and prime-influence information for every participant on all trials and enable performance on the influence aware trials to be directly compared to performance on the non-influence aware trials, within rather than between participants. With this information one can then quantify what contribution influence aware vs. non-influence aware trials make to AMP effects.

With this in mind, we designed a variant of the task known as the Influence-Awareness AMP or IA-AMP to investigate the following questions. First, would we observe an AMP effect for generic valenced primes? Second, and at the trial-by-trial level, are those effects moderated by a subset of trials, namely trials in which a participant is aware of the prime's influence on their evaluations? Third, are AMP effects at the group level moderated by inter-individual differences in awareness of influence of the primes? Fourth, does the "on-line" measure of awareness provided by the IA-AMP correlate with the post-hoc self-report awareness measures typically used in this literature? Finally, does influence awareness on the IA-AMP predict AMP effects better than post-hoc self-report measures?

Method

Sample Selection Strategy

Based on power analyses (i.e., 95% power to observe a medium effect size [$f^2 = 0.15$] in a linear regression analysis with a single predictor at the 0.05 alpha level), our *a priori* sample size after exclusions was 150 participants. Our sampling strategy involved recruiting 150 participants and then excluding those with incomplete data or who failed to meet inclusion criteria. Recruitment continued in batches of 10 until analyzable data was available for at least 150 participants.

Participants and Design

In total 214 participants took part, and of those, 147 (82 female) ranging in age from 18 to 65 years (*M* age = 34.9, *SD* = 11.7) provided complete and analyzable data. A single factor (*Prime Valence*: positive vs. negative) was manipulated within participants, and two dependent variables were assessed: influence awareness ratings and target stimulus evaluations on each trial.

Materials

The IA-AMP consisted of 12 positively and 12 negatively valence IAPS images (primes), a random selection of 120 of 200 possible Chinese pictographs (targets), and a white noise image (mask).

Procedure

Participants first provided informed consent and demographic information, and then completed an IA-AMP. Thereafter they completed a post-hoc self-report measure of awareness and exploratory questions.

IA-AMP. Prior to the task participants received a similar set of instructions as in the standard AMP. They were also told that the first image can sometimes bias their judgements of the Chinese pictographs and to try their absolute best not to let this happen. After each trial they

would be able to indicate if their judgement of the pictograph had been influenced by the first image by pressing the spacebar. If not, then they simply had to wait for the following trial.

The task consisted of 10 practice trials followed by 120 critical trials with similar parameters to the standard AMP. However, at the end of each trial, participants were given the opportunity to press the spacebar to indicate that their evaluation had been influenced by the prime on that trial. Specifically, a cue to "press spacebar if you felt you were influenced by the picture" was presented after each evaluation was made. This cue remained onscreen for 2000ms during which the spacebar could be pressed, followed by a 200ms inter-trial interval and the next trial (note: this response window was fixed regardless of whether a response was emitted or not).

Self-Report Awareness Measure. This measure was identical to that used in Payne et al. (2013; Experiment 1) and asked participants: "to what extent were your ratings of the Chinese symbols influenced by the pictures that appeared immediately before those symbols?". They could respond using on a 7-point Likert scale ranging from "Never" to "Almost always".

Exploratory Questions. A number of questions were asked concerning the intentionality of prime influence and contingency memory. These items were exploratory in nature were not part of our preregistered analyses.

Results

Data Preparation

For analyses at the trial-by-trial level we computed IA-AMP effects using the standard method outlined in Experiment 1. Given our interest in the *magnitude* of IA-AMP effects, regardless of their directionality, all analyses on trial-by-trial level effects examined absolute values (i.e., the difference in evaluations between the prime types, agnostic to the direction of the effect). We also calculated influence rates for each participant in the IA-AMP by dividing the

number of trials where they reported being influenced by the prime (i.e., by pressing the spacebar) by the total number of trials in the IA-AMP. [5]

Analytic Strategy

A logistic mixed-effects model was used to investigate our first question (is there evidence for an AMP effect) and second question (at the trial-by-trial level, is prime-consistent evaluation moderated by influence awareness). We opted for mixed-effects models given that they offer superior statistical power compared to more commonly used fixed-effects alternatives. To address our third question (are AMP effects at the group level moderated by those participants who are more highly aware of the prime's influence) we scored IA-AMP effects for each participant (*see below*) and entered these into linear regression models. To compare online (IA-AMP) and offline (self-report) measures of influence awareness, correlational and regression analyses were used.

Hypothesis Testing

Was There Evidence For An AMP Effect? A logistic mixed-effects model was carried out, with Valence Ratings (pleasant or unpleasant) of the target stimulus on each trial as the dependent variable, Prime Valence (pleasant or unpleasant) as the independent variable, and Participant as a random effect. This model acknowledges the non-independence of the multiple data points provided by each participant (i.e., the hierarchical nature of the data). Overall, an AMP effect emerged, such that participants were more likely to rate the target stimulus as positive when the prime valence was positive compared to when the prime valence was negative (and vice versa), OR = 3.23, 95% CI [3.02, 3.46], $p < .001$.

[5] In our preregistration we incorrectly specified this as the 'proportion of influenced to non-influenced trials' (i.e., a ratio rather than a proportion). We opted to treat influenced trials as a proportion of the total number of trials.

Are AMP Effects Moderated By Influence Awareness At The Trial Level? We extended the above model by adding influence awareness on each trial (influence aware vs. non-influence aware) as a fixed effect. This allowed us to determine if the relationship between Valence Rating and Prime Valence was moderated by that subset of influence aware trials. Results revealed an interaction between Prime Valence and influence awareness, such that AMP effects were far stronger on influence aware trials, OR = 15.61, 95% CI [13.17, 18.49], $p < .001$.

Are AMP Effects Moderated By Those Participants Who Are More Influence Aware? We then sought to determine if AMP effects were moderated by those participants who were more frequently aware of the prime's influence on their evaluations (i.e., whether awareness rates varied between individuals and whether this variation was associated with the magnitude of the AMP effect). An 'awareness rate' score was calculated for each participant by dividing the number of 'aware' trials by the total number of trials completed (i.e., 120). A linear regression analysis with AMP effect size as the dependent variable and influence awareness rate as a predictor variable was then conducted. Results indicated that influence awareness rate was a significant predictor of AMP effect size, $B = 0.42$, 95% CI [0.31, 0.53], $\beta = 0.53$, 95% CI [0.39, 0.57], $p < .001$.

Do Online (IA-AMP) And Post-Hoc (Self-Report) Measures Of Awareness Correlate With One Another? Simple correlations revealed that the IA-AMP and post-hoc awareness measures strongly associated with one another, $B = 0.14$, 95% CI [0.13, 0.16], $\beta = 0.83$, 95% CI [0.73, 0.92], $p < .001$.

Does Influence Awareness On The IA-AMP Predict AMP Effects Better Than Post-Hoc Self-Report Measures? Regression analyses were once again conducted with the two awareness measures added into the model. This allowed us to determine their relative

contribution in predicting AMP effects. Results indicated that only awareness assessed during the IA-AMP task predicted AMP effect sizes, $B = 0.38$, 95% CI [0.15, 0.61], $\beta = 0.42$, 95% CI [0.17, 0.67], $p = .001$; whereas awareness assessed after the task (post-hoc self-report) did not, $B = 0.02$, 95% CI [-0.02, 0.06], $\beta = 0.12$, 95% CI [-0.13, 0.37], $p = .341$). Comparison of the beta estimates' confidence intervals indicated that assessing for awareness during the task was a significantly better predictor of effects than doing so afterwards.

Discussion

The results from Experiment 2 are in-line with our preregistered hypotheses. AMP effects emerged and were moderated *at the trial-by-trial level* by performance on a subset of trials – namely – those where a participant was aware of the prime's influence on their evaluations. *At the group level,* AMP effects were moderated by participants who were highly influence aware and were significantly larger when calculated on the basis of the influence aware trials compared to when calculated on the basis of the non-influence aware trials. The online measure of awareness was a superior predictor of AMP effects than the offline measure.

Experiment 3: Awareness Assessed During an IA-AMP Predicts The Magnitude of Effects on a Previously-Completed Standard AMP (in the Same Attitude Domain)

One question that comes to mind is how performance on the IA-AMP relates to performance on a standard AMP. It may be that asking about influence on every trial serves to artificially raise awareness of the prime as well as its influence on evaluations. This in turn may lead to a stronger relationship between awareness and AMP effects than would normally occur in the standard version of the task. Therefore, in Experiment 3 we sought to not only replicate our previous findings but also address this new question. Participants were asked to complete a standard AMP with generic valenced primes followed by an IA-AMP with similar stimuli (i.e.,

both task variants indexed attitudes from the same domain). This approach provided us with a baseline AMP effect for each participant that was unperturbed by awareness probes as well as a separate influence awareness measure from that same person. If influence awareness rates on an IA-AMP completed at Time 2 correlate with standard AMP effects obtained at Time 1 then this would suggest that influence awareness may also be central to the standard AMP as well.

In short, Experiment 3 had both confirmatory and exploratory goals (all of which were preregistered). On the one hand we sought to confirm our earlier findings (i.e., that AMP effects would emerge on the IA-AMP; that these effects would be moderated by performance on the influence aware trials within a given individual, while at the group level, be moderated by highly aware participants). On the other hand, we also set out to explore the aforementioned question - would effects in the standard AMP be predicted by influence awareness rates in the IA-AMP? Based on our previous findings we predicted they would be.

We then explored a third and final question: would there be a difference in the magnitude of standard AMP effects relative to IA-AMP effects that are exclusively comprised of 'non-influence aware' trials? Recall that Payne et al. (2013; Experiment 3) asked a similar question when they compared scores on a standard AMP to those on a 'skip' AMP which was argued to provide a non-influence aware measure of evaluations. Unlike those authors (who found no difference between the two measures) we predicted that standard AMP effects would be significantly larger than those obtained from an IA-AMP comprised of exclusively "non-influence aware" trials, thus providing further support for the idea that influence awareness plays a key role in standard AMP effects as well.

Method

Sample Selection Strategy

Based on power analyses using identical criteria as Experiment 2, our *a priori* required sample size after exclusions was 150 participants. [6]

Participants and Design

206 participants took part, and of those, 176 (102 women) ranging in age from 18 to 64 years (M = 33.60, SD = 11.45) provided complete and analyzable data. A 2 (*Task Type*; standard vs. IA-AMP) x 2 (*Prime Type*; positive vs. negative) design was employed with both factors manipulated within participants. Two dependent variables were assessed: target stimulus evaluations and influence awareness responses.

Materials

AMP stimuli were similar to those used in Experiments 1 and 2.

Procedure

Participants first provided informed content and demographic information, and then completed a standard AMP, IA-AMP, the post-hoc self-reported awareness measure, and exploratory questions.

AMPs. Two version of the task were employed in Experiment 3: a standard AMP (similar to that used in Experiment 1) and an IA-AMP (similar to that used in Experiment 2). Both consisted of 72 trials and contained generic valenced prime stimuli.

Results

Analytic Strategy

A linear regression model was used to examine our confirmatory questions (i.e., if AMP effects would emerge on the IA-AMP; if these effects would be moderated by performance on

[6] We should note that more participants were sampled than originally specified in our preregistration due to an error in how exclusions were originally implemented in our data processing R script. Data collection was stopped when we believed we had 150 participants, as per the preregistration. A code review revealed that some participants had been erroneously excluded. The final analytic sample therefore includes these participants.

the influence awareness trials within a given individual, and be moderated by those participants who are more influence aware at the group level). A similar model was used to examine our first exploratory question (i.e., if influence awareness rates on the IA-AMP predict effect sizes in a previously completed standard AMP). A paired-samples *t*-test was used to investigate our second exploratory question (i.e., for differences between standard AMP effect sizes and non-influence aware only AMP effect sizes).

Data Preparation

Three AMP scores were calculated for each participant: an overall effect for the standard task, an overall effect for the IA-AMP, and a 'non-influence aware' effect based on those trials from the IA-AMP where participants did not press the spacebar (i.e., did not indicate awareness of the prime and its influence on their evaluations). This score notionally reflects an AMP effect based exclusively on non-influenced trials.

Hypothesis Testing

Was There Evidence For An AMP Effect And Was This Effect Moderated By Influence Awareness Within Individuals And At The Group Level? A significant effect emerged in both the standard AMP (OR = 3.10, 95% CI [2.87, 3.35], $p < .001$) and IA-AMP (OR = 4.66, 95% CI [4.30, 5.05], $p < .001$). At the trial-by-trial level, IA-AMP effects were moderated by influence aware trials, OR = 20.65, 95% CI [17.10, 24.94], $p < .001$. At the group level, IA-AMP effects were predicted by the influence awareness rates of participants, $B = 0.44$, 95% CI [0.34, 0.54], $\beta = 0.56$, 95% CI [0.44, 0.68], $p < .001$.

Does Influence Awareness On An IA-AMP Completed At Time 2 Predict Standard AMP Effects At Time 1? A regression analysis was conducted with standard AMP effect sizes as a dependent variable and influence awareness rate in the IA-AMP as a predictor variable.

Results indicated that influence awareness rates in the IA-AMP predicted the magnitude of effects in the standard AMP, $B = 0.34$, 95% CI [0.24, 0.45], $\beta = 0.44$, 95% CI [0.30, 0.57], $p < .001$.

Is There A Difference In The Magnitude Of Standard AMP Effects And Those Based Exclusively On Non-Influenced Trials? Results from a partially overlapping *t*-test (see Experiment 1) indicated that AMP effects based exclusively on non-influenced trials were significantly smaller than effects in the standard AMP, $t(163.85) = 5.09$, $p < .001$, $M_{diff} = 0.14$. A between-subjects Cohen's *d* was also calculated, although this should be interpreted with caution as it does not acknowledge the partial dependence among the data, $d = 0.41$, 95% CI [0.19, 0.63].

Discussion

Our findings replicated: AMP effects emerged and were moderated at the trial-by-trial level by performance on a subset of (influenced) trials, and at the group level by highly influence aware participants. We also extended these findings by showing that influence awareness during an IA-AMP at Time 2 predicted the size of standard AMP effects completed at Time 1. This suggests that asking about influence awareness of a trial-by-trial basis does not artificially raise awareness of the prime and its influence on evaluations. If it did then we would have expected no relationship between influence awareness rates and standard AMP effects to emerge, especially given that the IA-AMP was completed *after* the standard AMP. Yet influence awareness rates were strongly predictive of standard AMP effects, suggesting that people may in fact be aware of the prime, and use that stimulus when forming an evaluation of the target.

Finally, we obtained further evidence that conflicts with Payne et al.'s (2013; Experiment 3) claim that standard and non-influenced AMP effects do not differ from one another. Results indicated that AMP effects exclusively generated from non-influenced trials were significantly

smaller than standard AMP effects. These findings are consistent with what we observed in our direct replication attempt in Experiment 1 and converge on the same conclusion: AMP effects are stronger when participants are aware of the prime and its influence on evaluations.

Experiment 4: Awareness Assessed During an IA-AMP Predicts the Magnitude of Effects on a Previously-Completed Standard AMP (assessing Different Attitude Domains)

Experiment 4 represents an even stronger test of our claims. Imagine if participants first complete a standard AMP in one domain (political attitudes) and then complete an IA-AMP in a completely different domain (attitudes towards generic valenced stimuli). Now imagine if the same pattern of findings once again emerges. This would mean that a given participant's influence awareness rates at Time 2 in one domain would be predicting the magnitude of their AMP effects at Time 1 in an entirely different domain. If so, then this would provide even stronger evidence that influence awareness is stable within individuals and moderates their task performance across AMPs in different attitude domains.

With this in mind, we employed a similar design to Experiment 3 but with one simple change: we varied the attitudes domains being assessed by the standard AMP (political attitudes towards Donald Trump vs. Barack Obama) and the IA-AMP (attitudes towards generic valenced stimuli as in Experiments 1-3). If influence awareness rates reflect a stable (within-participant) pattern of responding regardless of content domain (politics vs. generic valenced primes), then influence awareness rates in a positive/negative IA-AMP at Time 2 should still predict effect sizes within a standard political AMP completed at Time 1.

Method

Sample Selection Strategy

Power analyses criteria were identical to Experiments 2 and 3 with an *a priori* required sample size after exclusions of 150 participants.

Participants and Design

Given that we were interested in assessing political attitudes we recruited a sample of residents from the USA who politically identified as Democrats. 175 participants took part in the study with data from 142 (74 women) ranging in age from 18 to 62 years (M = 31.90, SD = 10.41) eligible for analysis. The design, independent, and dependent variables were similar to those in Experiment 3.

Materials

The IA-AMP was identical to that used in Experiment 3. The standard AMP was also similar with the exception of the prime stimuli which now 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).

Procedure

The procedure was similar to Experiment 3 with two changes: the standard AMP now assessed political attitudes and the addition of two new exploratory questions.

Exploratory Questions. In addition to the same exploratory questions asked in Experiments 1-3 we also assessed for demand compliance and political alignment.

Results

Analytic Strategy

For this and all subsequent experiments, we divide results into two sections. ‘Replication hypotheses’ refer to hypotheses that were first made in one of our previous experiments and which were retested in the current experiment. These will be briefly reported given that a

detailed treatment is provided in a previous experiment. 'Critical hypotheses' refer to new hypotheses being made within a given experiment.

Data Preparation

Data preparation was similar to that of Experiment 3.

Hypothesis Testing

Replicated Hypotheses: Was There Evidence For An AMP Effect And Was This Effect Moderated By Influence Awareness Within Individuals And At The Group Level? A significant effect emerged on the IA-AMP, OR = 3.09, 95% CI [2.84, 3.37], $p < .001$, and standard AMP, OR = 3.85, 95% CI [3.53, 4.19], $p < .001$. At the trial-by-trial level, effects were moderated by the subset of trials where a participant was influence aware, OR = 40.83, 95% CI [31.98, 52.13], $p < .001$; while at the group level, effect sizes were moderated by inter-individual differences in influence awareness, both within the IA-AMP ($B = 0.57$, 95% CI [0.40, 0.74], $\beta = 0.49$, 95% CI [.34, 0.63], $p < .001$) and for a previously completed AMP ($B = 0.54$, 95% CI [0.44, 0.65], $\beta = 0.65$, 95% CI [0.53, 0.78], $p < .001$).

Critical Hypotheses: Does Influence Awareness On A Generic Valence IA-AMP Completed At Time 2 Predict Political AMP Effects At Time 1? A regression analysis with awareness rate in the IA-AMP as a predictor and standard AMP effect size as a dependent variable revealed that the former significantly predicted the latter, $B = 0.54$, 95% CI [0.44, 0.65], $\beta = 0.65$, 95% CI [0.53, 0.78], $p < .001$.

Discussion

Our prior findings once again replicated: AMP effects emerged and were moderated at the trial-by-trial level by performance on influence aware trials, while at the group level, they were moderated by participants scoring high in influence awareness. Not only did influence

awareness assessed by an IA-AMP retrospectively predict standard AMP effects, but it did so even when these tasks were assessing attitudes in entirely different domains. Taken together, Experiments 3 and 4 suggest that influence awareness is stable within individuals and moderates the magnitude of the AMP effect, even between attitude domains. It may be the case that, at least at the group level, the majority of the variance in AMP effects does not represent a measure of implicit evaluations *in general*, but rather the evaluations of a subset of individuals who are highly influence aware. We will return to this idea below.

Experiment 5: The AMP's Predictive Utility is Based on Influence Aware Trials

Experiments 2-4 indicate that a subset of influence aware trials strongly moderate AMP effects at the trial-by-trial level, and inter-individual differences in influence aware rates moderate the magnitude of the AMP effect. Influence awareness rates on one AMP predict how one will respond on another, and this is true when both tasks assess the same or different attitude domains.

In Experiment 5 we set out to further replicate our findings while addressing three new questions. On the one hand, we wanted to know if influence awareness also played a key role in the AMP's predictive utility (i.e., its ability to discriminate between two known-groups). We therefore recruited two groups of participants (Democrats and Republicans) and asked them to complete a political IA-AMP followed by a positive/negative IA-AMP. Our preregistered hypothesis was that the AMP's ability to predict whether a person was a Democrat or Republican would be higher when effects were solely derived from influence aware trials and lower when they were derived from non-influence aware trials.

At the same time, we wanted to know if there was intra-individual stability in influence awareness from one AMP to another. Experiments 2-3 offered indirect evidence such that

influence awareness rates in the IA-AMP predicted the same person's scores in the standard AMP. However, a more direct demonstration requires that we have a measure of awareness on both tasks. We therefore examined if a participant's influence rate on one IA-AMP was correlated with their influence rate on a second IA-AMP.

Finally, a unidirectional relationship between influence awareness and AMP effect sizes emerged in Experiments 2-4. However, given that we were now administering two IA-AMPs, we could also now test for a bidirectional relationship between these variables (i.e., if influence rates in AMP #1 predict effects in AMP #2, and if influence rates from AMP #2 predict effects in AMP #1). Demonstrating such a relationship would suggest that *influence rates in general* are predictive of *AMP effects in general,* yet further evidence supporting the idea that AMP effects are highly dependent on awareness of the primes and its influence on one's evaluations.

Method

Sample Selection Strategy

Power analyses for interactions in mixed-effects models are difficult to determine due to the large increase in the number of parameters involved, therefore no power analysis was conducted for our first analysis. For our second analysis, we used the pwr package in R to compute the number of participants required to detect a medium f^2 effect size (i.e., 0.15) in a regression analysis with a single IV, at the conventional alpha level (.05) and at 95% power. Given these criteria, 89 participants would be required. The aforementioned power analysis is also applicable for our third analysis. With 89 participants, at a standard alpha level and a power of .90, we would be able to detect a correlation of $r = .33$. We chose to collect data from at least 200 participants (100 Democrats and 100 Republicans) based on the availability of resources.

Participants and Design

A total of 334 participants took part, and of these, 207 (105 Democrats, 102 Republicans; 106 women) ranging in age from 18 to 65 years (M = 34.03, SD = 11.15) provided complete and analyzable data. A 2 (*Content Type*; political vs. generic valence) x 2 (*Prime Type*; positive vs. negative) x 2 (*Political Orientation*; Democrat vs. Republican) design was employed with the first two factors manipulated within and the third manipulated between participants. Influence awareness rates and evaluations were the two dependent variables.

Materials

Two IA-AMPs were employed. The first was an IA-AMP that used the same political primes as in Experiment 4 while the second was a generic valence IA-AMP identical to that used in our previous studies.

Procedure

Participants provided informed consent and demographic information. They then completed a politics IA-AMP, a generic valence IA-AMP, the post-hoc self-reported awareness measure, and exploratory questions.

Results

Analytic Strategy

Our replicated hypotheses were assessed using a similar analytic strategy as before. To examine the relationship between influence awareness and the AMP's predictive validity, two AMP scores were calculated, one based solely on the influence aware trials and another based solely on the non-influence aware trials. We then used two between-groups *t*-tests to examine their relative ability to discriminate between Democrats and Republicans. To examine the consistency of influence awareness rates within participants across different AMPs, a simple correlation test was used. Finally, two linear regression models were used to assess the

bidirectional relationship between influence rates and AMP scores. In the first regression, influence awareness rate was taken from the politics IA-AMP, and the effect size from the positive-negative IA-AMP. In the second regression, influence awareness rate was taken from the positive-negative IA-AMP and effect sizes from the politics IA-AMP. [7]

Hypothesis Testing

Replicated Hypotheses: Was There Evidence For An AMP Effect And Was This Effect Moderated By Influence Awareness Within Individuals And At The Group Level? A significant effect emerged on the positive-negative IA-AMP, OR = 3.27, 95% CI [3.05, 3.51], $p < .001$. Effects also emerged on the political IA-AMP and in the opposite direction for Republicans and Democrats. Specifically, a significant interaction effect was obtained for Prime Type (Trump vs. Obama) and Political Orientation (Republicans vs. Democrats), OR = 0.11, 95% CI [0.10, 0.13], $p < .001$. At the trial-by-trial level, effects on both IA-AMPs were moderated by the subset of trials where a participant was influence aware (valence: OR = 29.14, 95% CI [23.72, 35.79], $p < .001$; politics: OR = 197.70, 95% CI [131.65 296.91], $p < .001$) while at the group level, effect sizes on a given IA-AMP were moderated by inter-individual differences in influence awareness on that task (valence: $B = 0.49$, 95% CI [0.40, 0.58], $\beta = 0.61$, 95% CI [0.50, 0.72], $p < .001$; politics: $B = 0.63$, 95% CI [0.53, 0.74], $\beta = 0.64$, 95% CI [0.54, 0.75], $p < .001$).

Critical Hypotheses: Does Influence Awareness Moderate The AMP's Predictive Validity? Results indicated that IA-AMP effects based solely on influence aware trials were superior in discriminating between Democrats and Republicans ($d = 2.08$, 95% CI [1.62, 2.55])

[7] Note that in Experiments 2-4 we focused on the absolute magnitude of AMP effect sizes. Here in Experiment 5 we took the directionality of AMP effects into account when comparing the AMP scores of Republicans to those of Democrats. In all other cases, absolute AMP scores were assessed when testing hypotheses at the participant level.

than effects based solely on the non-influence aware trials (d = 0.62, 95% CI [0.33, 0.91]), Q(df = 1) = 27.51, p < .001.[8] As shown in Figure 2, discriminability between the known-groups was primarily moderated by those trials where people are aware of the prime's influence on their evaluations.

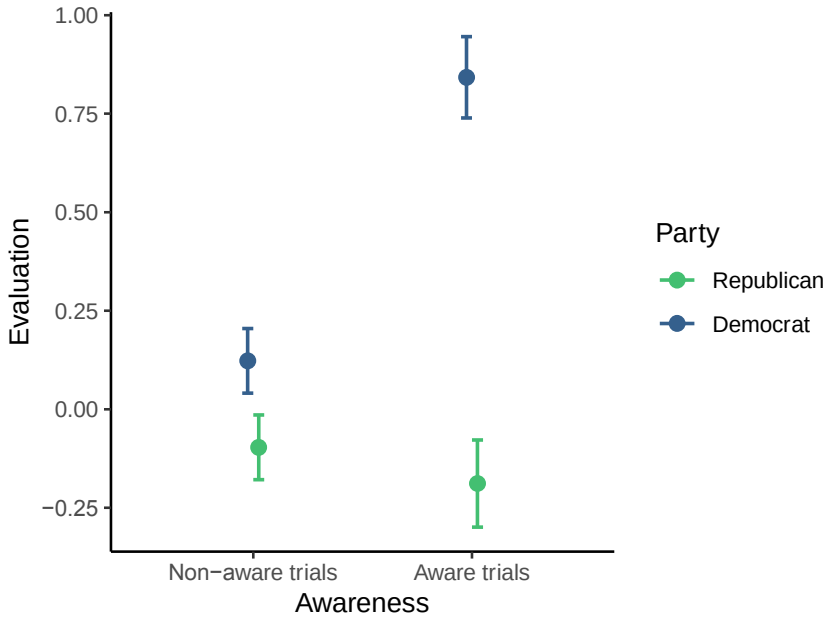

Figure 2. The political IA-AMP's ability to discriminate between Democrats and Republicans on the basis of influence aware and non-influence aware trials. A negative score indicates a preference for Trump over Obama whereas a positive score indicates a preference for Obama over Trump. Error bars represent 95% confidence intervals.

Are Influence Awareness Rates Consistent Within Individuals And Across Different AMPs? Results revealed a strong correlation between influence awareness rates in the two task variants, r = 0.82, 95% CI [0.77, 0.86], p < .001.

[8] Our preregistration stated that we would compare differences between these conditions via the confidence intervals on the two Cohen's d estimates. We subsequently discovered a method to produce a p value for this comparison via the metafor package's heterogeneity test. Both are reported and results are congruent across these analytic choices.

Does Influence Awareness In One IA-AMP Predict Performance In Another IA-AMP, And Is This Relationship Bidirectional? Results indicated that influence awareness rates in the politics IA-AMP predicted scores in the positive-negative IA-AMP, $B = 0.46$, 95% CI [0.36, 0.55], $\beta = 0.54$, 95% CI [0.42, 0.66], $p < .001$, and that influence awareness rates in the positive-negative IA-AMP predicted scores in the politics IA-AMP, $B = 0.49$, 95% CI [0.38, 0.60], $\beta = 0.52$, 95% CI [0.40, 0.63], $p < .001$. It is also useful to consider the implications of these results in terms of temporal order rather than domain. Although it was not part of our original research plan, these results also suggest that the temporal order of the tasks, and therefore the order of assessment of the AMP effect versus the influence rate, does not matter. Participants always completed the politics IA-AMP first and the valence IA-AMP second. The influence rate in the politics IA-AMP (completed first) predicted the absolute magnitude of the valence IA-AMP (completed second), B = 0.46, 95% CI [0.36, 0.55]. Equally, the influence rate of the valence IA-AMP (completed second) predicted (or more accurately 'postdicted') the absolute magnitude of the politics IA-AMP (completed first), B = 0.49, 95% CI [0.38, 0.50]. The very similar estimates and strongly overlapping confidence intervals provide no evidence that order of presentation moderated the effect.

Discussion

Experiment 5 offers three new insights into the relationship between influence awareness and AMP effects. First, the predictive validity of AMP effects based solely on influence aware trials is far superior to that of effects based solely on non-influence aware trials. Second, a given person's influence awareness rate on one AMP is strongly correlated with their influence awareness on another AMP, even when those tasks are targeting entirely different domains. Third, the relationship between influence awareness and AMP scores is *bidirectional*, insofar as

AMP effect sizes predict how influence aware one will later report being, and how influence aware one is at an earlier point in time will predict the later magnitude of their AMP effects.

In short, these findings provide yet further support for the idea that (a) AMP effects are produced by a subset of influence aware trials and participants who are highly influence aware, and (b) that the influence aware participants who are mainly responsible for AMP effects in one domain are the same participants who are responsible for effects in another domain. They also imply that the AMP's predictive validity is heavily dependent on influence awareness. Although non-influence aware trials retain some degree of predictive validity, this pales in comparison to that of influence aware trials.

Experiment 6: Performance on the Mann et al. AMP is Also Moderated by Influence Awareness

We are not the first to argue that AMP effects are strongly moderated by a well-defined subset of participants. In a recent review of the AMP literature, Mann et al. (2019) noted that data from AMP studies exhibit a strong bimodal distribution, with a subset of participants showing a very strong AMP effect, and the others producing scores that follow a normal distribution (for related arguments also see Bar-Anan & Nosek, 2012).

Mann et al. argued that this cluster of extreme scoring participants (i.e., those who responsible for the bimodality) represent a small group of *intentional* responders, whereas the remaining participants reflect *unintentional* responders. They sought to eliminate the contaminating influence of these intentional responders (and thus reduce this bimodality) by creating a new and improved variant of the AMP. This task employed visually stimulating paintings as target stimuli, rather than the less visually stimulating Chinese pictographs, in order to increase the chances that participants would pay attention to the target rather than prime. They

also included additional instructions imploring participants to avoid intentionally responding to the prime while reassuring them that it was acceptable if they sometimes did so. Mann et al. concluded that their modifications to the AMP decreased bimodality compared to a standard AMP (and thus reflected a less intentional measure of evaluations).

In Experiment 6 we examined if awareness of influence of the prime is also reduced in the Mann et al. AMP (referred to hereafter as the 'Mann AMP'). We did so by replicating Experiment 3 using this new version of the task. That is, participants were first asked to complete a standard Mann AMP and then complete a version of that task where they could also indicate if they were aware of the prime and its influence on their evaluations (referred to as the 'Mann IA-AMP'). If the Mann AMP successfully limits or excludes influence aware trials and participants, then we should not expect to replicate our prior findings. However, if we do replicate those findings, then it would suggest that even this purportedly 'improved' version of the task is also heavily dependent on influence awareness.

Based on our findings to date, we preregistered two hypotheses. On the one hand, we argued that, at both the trial- and trial-by-trial level, Mann IA-AMP effects would be heavily moderated by influence awareness. On the other hand, we hypothesized that influence awareness rates of a given participant in the Mann IA-AMP at Time 2 would predict the size of that same person's Mann AMP effects completed at Time 1.

Method

Sample Selection Strategy

Power analyses began with an examination of the association between the IA-AMP influence awareness rates and absolute AMP effects observed in Experiment 2. Results from that study indicated that this association was in the range $\beta = 0.56$, 95% CI [0.44, 0.68]. However, we

were unsure whether the Mann et al. modification to the AMP would impact the magnitude of this association compared to our previous studies. We therefore opted to power our analyses to detect an even smaller effect size (i.e., $\beta = .20$). To power a regression analysis to detect a $\beta = .20$ at a 0.05 alpha level (two-tailed) with 95% power requires 320 participants. This was defined as our *a priori* sample size after exclusions and participants were sampled in a similar fashion to our previous experiments.

Participants

410 participants took part, and of those, 330 (171 women) ranging in age from 18 to 65 ($M = 33.40$, $SD = 11.05$) provided complete and analyzable data.

Procedure

A similar procedure to Experiment 2 was used with one exception: the standard AMP was replaced with Mann et al.'s AMP, and the IA-AMP was replaced with a Mann et al. variant of our IA-AMP.

AMPs. In line with Mann et al. (2019), each AMP consisted of 10 practice trials, 60 main trials, 12 positive and 12 negative valence images, and 60 paintings. All parameters of these AMPs were identical to the AMP of Mann et al., with one exception: rather than use face images as prime stimuli, we used generic valenced IAPS images (identical to those used in Experiment 2).

Results

Analytic Strategy

Our analytic strategy was identical to that of Experiment 2.

Data Preparation

Our data preparation was identical to that of Experiment 2.

Hypothesis Testing

Replication Hypotheses: Do We Find Evidence For Mann IA-AMP Effects? A significant effect emerged on both the Mann AMP, OR = 3.72, 95% CI [3.48, 3.98], $p < .001$, and the Mann IA-AMP, OR = 4.36, 95% CI [4.08, 4.67], $p < .001$.

Critical Hypotheses: Does Influence Awareness Predict Mann IA-AMP Effects At The Trial Level And Trial-By-Trial Level? Results revealed an interaction between influence awareness and Prime Type in the Mann IA-AMP, OR = 16.30, 95% CI [13.79, 19.28], $p < .001$, such that IA-AMP effects were moderated by influence aware trials. Results also indicated that influence awareness rates significantly predicted the magnitude of Mann IA-AMP effects, B = 0.54, 95% CI [0.47, 0.62], $\beta = 0.61$, 95% CI [0.53, 0.70], $p < .001$.

Does Influence Awareness On A Mann IA-AMP Completed At Time 2 Predict The Magnitude Of Mann AMP Effects Completed At Time 1? Results indicated that influence awareness rates in the Mann IA-AMP predicted scores in the previously completed Mann AMP, $B = .38$, 95% CI [0.30, 0.47], $\beta = .42$, 95% CI [0.32, 0.52], $p < .001$.

Non Preregistered Analyses: Does The Predictive Utility Of Influence Awareness Vary Between The Standard And Mann AMPs? Following data collection, we noted that effect sizes in the influence rates predicting Mann AMP effects was relatively similar to that reported in Experiment 2. We therefore examined if effect sizes for this analysis in Experiment 2 (where a standard AMP was used) differed significantly from those in Experiment 6 (where a Mann AMP was used). Data from Experiments 2 and 6 were pooled and a similar regression model was constructed as used in those experiments (i.e., Influence Rate as IV, [Mann] AMP effect as DV), also adding AMP type (i.e., Experiment) as a fixed effect in the model. If Influence Rate significantly differed in how well it predicted AMP effects between the standard

and Mann AMPs, then an interaction between Influence Rate and AMP type (i.e., Experiment) should emerge. However, no such interaction was observed, $B = 0.04$, 95% CI [-0.09, 0.18], $\beta = 0.05$, 95% CI [-0.11, 0.21], $p = .534$. In order to quantify evidence for the absence of this interaction, a Bayes Factor for the interaction effect was computed using the BayesFactor R package (Morey & Rouder, 2019) by comparing models within and without this interaction effect. This Bayesian analysis using the default prior (Cauchy distribution placed on the effect size with scaling factor $r = 0.5$) revealed moderate evidence in support of the null hypothesis, $BF_{10} = 0.12$.

Discussion

The Mann et al. AMP was recently introduced with the aim of eliminating a similar phenomenon as in our previous experiments: namely, that only a subset of participants contribute to the AMP effect. However, we found the same pattern in that version of the task as we did in the standard task: a subset of influence aware trials (within participants), and highly influence awareness participants (between participants) strongly moderated AMP effects. Influence awareness rates in the Mann IA-AMP also predicted effects sizes in a previously completed Mann AMP. Furthermore, the extent to which influence awareness rates predicted the size of AMP effect sizes did not differ from, and was credibly equivalent to, what was observed in Experiment 2 with the standard AMP. Put simply, we obtained the same pattern of outcomes as reported in Experiments 2-5 with a variant of the AMP specifically designed to eliminate subset effects seen in other AMP research.

Experiment 7: Prospective Influence Awareness Measures Also Predict AMP Effects

Experiments 2-6 show that influence awareness is a strong moderator of the magnitude of AMP effects at the individual and group levels, across different versions of the task (standard,

Mann et al., IA-AMP), and within the same or between different content domains. Critically, however, we always assessed influence awareness in a *retrospective* fashion such that people were first asked to emit an evaluative response and only then reflect on it. Although this reflection occurs mere milliseconds after the evaluative response itself, it is still in some sense post-hoc. We therefore wanted to know if our findings would replicate when a *prospective* measure was used, one where awareness is assessed before the evaluative response is emitted. With this in mind, we conducted an exact replication of Experiment 3 wherein a standard AMP was completed followed by an IA-AMP with primes from the same attitude domain. This IA-AMP was modified so that participants first signaled if they were influence aware and only then provided their evaluative response to the target. In this way, it would not be possible for participants to confabulate influence awareness based on their previously emitted evaluative response, because that response had not yet been emitted. If our findings were to replicate this would lend still further evidence to the idea that people are aware of the influence of the prime on evaluative responses, both in retrospective and prospective ways (see Figure 1).

Method

Sample Selection Strategy

Power analyses were identical to that of Experiment 3 and the sampling strategy was identical to previous experiments.

Participants

184 participants took part, and of those, 153 (94 women) ranging in age from 18 to 63 (M = 32.58, SD = 10.86) provided complete and analyzable data.

Procedure

The procedure was similar to that of Experiment 3 with one exception: the IA-AMP was changed from a retrospective to a prospective measure of influence awareness.

IA-AMP. This task consisted of the same parameters as in previous studies with one change: after the presentation of the target stimulus, but before emitting the evaluative response, participants were given the opportunity to press the spacebar to indicate if they believed their response to the target *will be influenced* by the prime. This was achieved through the presentation of the cue to "Press spacebar if the picture will influence your response to the Chinese symbol" for a fixed 2000ms interval. The above sentence was removed from the screen following a response, although the response window was fixed regardless of whether a response was emitted or not.

Results

Analytic Strategy

Our analytic strategy was similar to that of Experiment 3.

Data Preparation

Data preparation was identical to that of Experiment 3.

Hypothesis Testing

Replication Hypotheses. A significant effect emerged on both the standard AMP, OR = 2.21, 95% CI [2.04, 2.39], $p < .001$, and the prospective IA-AMP, OR = 2.69, 95% CI [2.48, 2.92], $p < .001$.

Critical Hypotheses: Are Prospective IA-AMP Effects Moderated By Influence Awareness? Influence awareness moderated evaluations at the trial-by-trial level, OR = 7.29, 95% CI [6.02, 8.83], $p < .001$, and inter-individual differences in influence awareness moderated

the magnitude of IA-AMP effects at the group level, $B = 0.54$, 95% CI [0.43, 0.64], $\beta = 0.63$, 95% CI [0.50, 0.75], $p < .001$.

Does Prospective Influence Awareness In The IA-AMP Predict Standard AMP Effects? A person's influence awareness rate in the prospective IA-AMP completed at Time 2 predicted the magnitude of their effect in a standard AMP completed at Time 1, $B = 0.41$, 95% CI [0.27, 0.54], $\beta = 0.45$, 95% CI [0.30, 0.59], $p < .001$.

Discussion

A prospective measure of influence awareness yielded similar findings to the retrospective measures used in Experiments 1-6. Thus a post-hoc confabulation account is a poor candidate for explaining these findings as well as the other outcomes reported thus far.

Experiment 8: Prospective Influence Awareness Measures (Prior to Prime Presentation) Also Predict AMP Effects

In our final study we wanted to replicate and extend our findings with the prospective measure even further. Experiment 7 assessed for influence awareness before an *overt* evaluative response was emitted. However, it is possible that people may still have formed a *covert* evaluation of the target after having seen the prime. If so, and if they also recognized that this covert evaluation was consistent with the valence of the prime when completing the influence awareness question, then this consistency may have formed the basis of a post-hoc confabulation regarding their awareness of the source of this covert evaluation. We therefore had participants register their influence awareness response *before* seeing the target stimulus at all (see Figure 1). In this way, they could not form a covert evaluation of the target stimulus, nor could their performance on the influence awareness measure be a post-hoc confabulation, because the target stimulus had not yet been presented. In such a situation (post-hoc) confabulation is not possible.

Method

Sample Selection Strategy

Power analyses were similar to Experiment 3 and the sampling strategy was similar to prior studies.

Participants

188 participants took part, and of those, 154 (89 women) ranging in age from 18 to 64 (M = 29.81, SD = 10.98) provided complete and analyzable data.

Materials

Materials were similar to Experiment 3 with the exception that the influence awareness response emitted on each trial was recorded prior to the presentation of the target.

Procedure

The procedure was similar to Experiment 3.

Results

Hypothesis Testing

Replication Hypotheses. A significant effect emerged on both the standard AMP, OR = 2.17, 95% CI [2.01, 2.35], $p < .001$, and the prospective IA-AMP, OR = 2.58, 95% CI [2.38, 2.80], $p < .001$.

Critical Hypotheses: Are Prospective IA-AMP Effects Moderated By Influence Awareness? At the trial level, prospective influence awareness moderated evaluative responses, OR = 6.26, 95% CI [5.21, 7.51], $p < .001$. This was also the case at the participant level: participants' prospective influence awareness rates strongly moderated the magnitude of their IA-AMP effect, B = 0.49, 95% CI [0.37, 0.61], β = 0.55, 95% CI [0.41, 0.68], $p < .001$.

Does Prospective Influence Awareness In An IA-AMP Completed At Time 2 Predict Effects In A Standard AMP Completed At Time 1? Influence awareness rates in the prospective IA-AMP significantly predicted the magnitude of their effect in a previously completed standard AMP, $B = 0.34$, 95% CI [0.20, 0.49], $\beta = 0.35$, 95% CI [0.20, 0.50], $p < .001$.

Discussion

Results indicate that a prospective measure of influence awareness administered between the prime and target stimulus, thereby removing the possibility of confabulation, resulted in the same pattern of results as in Experiments 1-7 (see Figures 3 and 4).

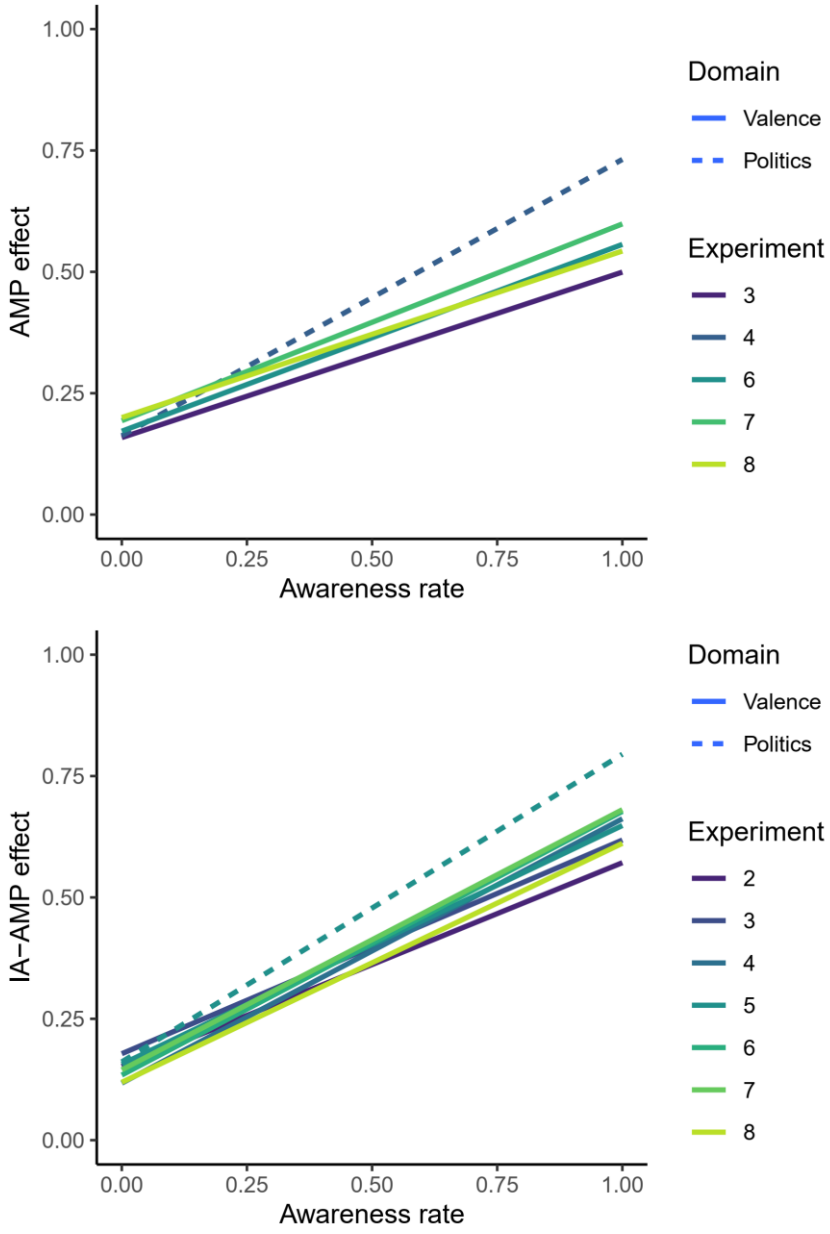

Figure 3. Influence awareness rates on the IA-AMP and the absolute magnitude of AMP effects on the IA-AMP (upper panel) and a previously completed standard AMP (lower panel), across Experiments 2-8.

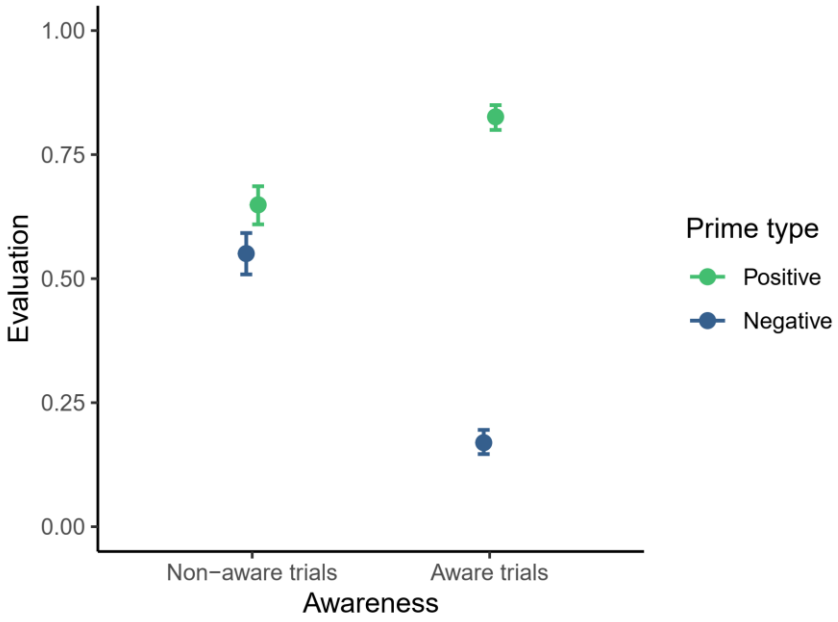

Figure 4. Trial-level influence awareness moderates the magnitude of IA-AMP effects. Point estimates represent marginal means from the meta-analytic model of Experiments 2-8 and their 95% Confidence Intervals.

Meta-Analyses

Meta-analyses were conducted in order to estimate the effect with greater precision across studies, and to estimate heterogeneity in the effect between experiments and across methodological variations. Meta-analyses were conducted using the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015). Meta-analyses were not preregistered, although the hypotheses assessed within them and their model specifications were identical to those preregistered in the original experiments, with the addition of a random intercept for experiment. Data from all novel experiments (2-8) were included (total N = 1309, k = 7).

The AMP Effect Is Strongly Moderated By Awareness

Inter-Individual Differences In Awareness Moderate The IA-AMP Effect

As in the preregistered analyses for the individual studies, we assessed whether the absolute magnitude of the AMP effect on the IA-AMP was associated with the influence awareness rate on that IA-AMP. Results demonstrated that a large proportion of the variance in AMP effects was attributable to the influence awareness rate between participants, $B = 0.52$, 95% CI [0.48, 0.55], $\beta = 0.60$, 95% CI [0.56, 0.64], $p < .001$.

Recall that the AMP effect is the difference in evaluations on trials involving positive versus negative primes, and can range from 0 (evaluations unrelated to prime valence) and 1 (all evaluations congruent with primes). The model intercept was $B = 0.14$, 95% CI [0.12, 0.16], $\beta = -0.01$, 95% CI [-0.07, 0.05], $p < .001$. At the two extremes, in participants who report being aware of the influence of the prime on their evaluations on 0% of trials, the estimated marginal mean AMP effect on the IA-AMP was therefore 0.14. In contrast, in participants who report being aware of the influence of the prime on their evaluations on 100% of trials, the estimated marginal mean AMP effect on the IA-AMP was 0.66. The AMP effect was therefore estimated to be three times larger in fully influence aware participants than fully non-influence aware participants. Very little of the variance was attributable to differences between experiments ($R^2 = .07$) compared to influence awareness ($R^2 = .36$). Given the methodological differences between the experiments, this implied that the strong moderation of the AMP effect on the IA-AMP by inter-individual differences influence awareness had good generalizability.

Inter-Individual Differences In Awareness On The IA-AMP Moderate The AMP Effect On A Previously Completed AMP

An identical set of analyses was conducted using the AMP effect on the standard AMP rather than the IA-AMP as the dependent variable. Results demonstrated that a large proportion

of the variance in AMP effects was attributable to the influence awareness rate between participants, $B = 0.39$, 95% CI [0.34, 0.44], $\beta = 0.42$, 95% CI [0.37, 0.48], $p < .001$.

The model intercept was $B = 0.18$, 95% CI [0.15, 0.21], $\beta = 0.01$, 95% CI [-0.08, 0.09], $p < .001$. At the two extremes, in participants who report being aware of the influence of the prime on their evaluations on 0% of trials, the estimated marginal mean AMP effect on the IA-AMP was 0.18. In participants who report being aware of the influence of the prime on their evaluations on 100% of trials, the estimated marginal mean AMP effect on the IA-AMP was 0.57. The AMP effect was therefore estimated to be three times larger in fully aware participants than fully non-aware participants. Very little of the variance was attributable to differences between experiments ($R^2 < .01$) compared to influence awareness ($R^2 = .18$). Given the methodological differences between the experiments, this again implied that the strong moderation of the AMP effect on a standard AMP by inter-individual differences influence awareness had good generalizability.

In summary, knowing an individual's influence awareness rate is sufficient to predict the magnitude of their AMP effect on a standard AMP that was completed prior to capturing the influence awareness rate (i.e., the AMP effect could not have been perturbed as awareness was only asked about later). This effect was found across 7 studies with very little evidence of heterogeneity, suggesting high replicability and generalizability across methodological variations (e.g., when within each IA-AMP trial influence awareness was assessed, and the domain being assessed; see Figure 3). As noted in Experiment 4, this effect holds even when the IA-AMP (used to capture the influence awareness rate) and the standard AMP are assessing different domains (valence vs. politics).

Trial-By-Trial Awareness Moderates The AMP Effect

IA-AMP effects were found to be moderated not only by inter-individual differences in awareness (as in the above analysis), but also intra-individual at the trial level, OR = 15.46, 95% CI [14.41, 16.59], $p < .001$ (see Figure 4).

What Is The Distribution Of Influence Awareness Across Participants?

Data were pooled in order to understand the distribution of influence awareness rates between participants. This analysis was exploratory and not included in the preregistrations for the individual experiments. Hartigan's dip test demonstrated non-unimodality ($D = 0.03$, $p < .001$). As a robustness test, we also analyzed the subset of AMPs that employed the Mann et al. (2019) modifications, which also demonstrated non-unimodality ($D = 0.03$, $p = .030$). Visual inspection of distribution kernel-density plots demonstrated clear bimodality (see Supplementary Materials). Gaussian kernel density estimation was used to estimate the two modes: influence awareness rates were found to cluster around participants being either fully non-aware (Mode = .01) or fully awareness (Mode = .97; range = 0 to 1). This provided convergent evidence that it is the subset of highly influence aware participants and their subset of influence aware trials that represent the majority of variance in observed AMP effects.

General Discussion

Over the past 15 years people have used the AMP under the assumption that its effects represent an implicit measure of attitudes, stereotypes, and other biases. Effects on the task are said to be 'implicit' insofar as prime stimuli influence how target stimuli are evaluated without a person's intention for this to happen, or awareness that such an occurrence has taken place. Although a number of papers have previously investigated the intentionality of AMP effects (e.g., Bar-Anan & Nosek, 2012; Payne et al., 2013; Gawronski & Ye, 2014), far less attention has been paid to the issue of awareness. Across eight preregistered, highly-powered studies,

including a direct replication and meta-analyses, we re-examined the implicitness of AMP effects in terms of whether participants are indeed unaware of the prime's influence on their evaluations. In what follows we summarize our findings and then discuss their implications for the AMP as well as the implicit and explicit accounts more generally.

Overview of Findings

Experiment 1 began with a replication attempt of Payne et al.'s (2013) work with the 'skip' AMP. The finding that 'skip' AMP effects (i.e., those where 'influenced' responses have been detected and removed) are no different to standard AMP effects is viewed as strong support for the implicit account. We examined if these findings would replicate using an improved and more highly-powered design. Results indicated that the original findings did not replicate, such that scores on the standard AMP were significantly larger than those on the skip AMP. Exploratory analyses also revealed that a given person's awareness of the prime's influence on their evaluations (during the skip AMP at Time 2) was strongly related to the magnitude of their effects in the standard AMP at Time 1.

A limitation of the 'skip' AMP is that it forces people to either skip or evaluate the target and thus only provides partial data. To overcome this we developed an influence aware IA-AMP in Experiment 2 wherein participants rated the target (thus providing evaluative information) and then indicated if that evaluation had been influenced by the prime (thus providing influence information). Results indicated that AMP effects emerged and that these were moderated at the trial-by-trial level by a subset of (influence aware) trials, and AMP effects were produced predominantly by highly influence aware participants at the group level.

In Experiments 3-4 we controlled for the possibility that by probing for influence awareness on each trial of the IA-AMP we artificially altered the relationship between awareness

and AMP effects. Participants now completed a standard AMP at Time 1 and an IA-AMP at Time 2, either from the same (Experiment 3) or different attitude domains (Experiment 4). Because the standard AMP was always completed prior to the IA-AMP, effects on the former were always unperturbed by modifications to the latter. Yet in both studies influence awareness during an IA-AMP at Time 2 predicted the magnitude of standard AMP effects at Time 1, indicating that influence awareness is a stable (within-participant) pattern of responding that holds within and between content domains.

Experiment 5 extended our analyses to three additional questions: is the predictive validity of the AMP effect also dependent on influence awareness; is there intra-individual stability in influence awareness from one AMP to another; and is the relationship between influence awareness and AMP effects bidirectional, such that the presence of one predicts the presence of the other. Two groups of participants (Democrats and Republicans) first completed a political IA-AMP and then an IA-AMP with generic valenced primes. We found that the AMP's ability to correctly classify a person as a Democrat or Republican was superior when effects were based solely on influence aware trials and inferior when based solely on non-influence aware trials. A given person's influence awareness rate on one AMP was also strongly correlated with their influence awareness on another AMP, even when those tasks target entirely different attitude domains. Finally, the predictive relationship between influence awareness and AMP scores was bidirectional; influence awareness from AMP #1 predicted the AMP effect of AMP #2, and vice versa.

Experiment 6 took a newly developed version of the AMP that purportedly reduces subset effects within the AMP (the Mann et al. [2019] AMP) and examined if influence awareness also plays a role here too. Participants first completed a standard Mann AMP and then

a Mann IA-AMP. Once again, the same pattern of findings emerged: a subset of influence aware trials (within participants), and highly influence awareness participants (between participants) were responsible for AMP effects. Influence awareness rates in the Mann IA-AMP also predicted effects sizes in a previously completed Mann AMP. Put simply, the same pattern of outcomes emerged even within a variant of the task designed to optimize the implicitness of the AMP.

In our final two studies we modified the IA-AMP so that influence awareness was measured *prospectively*, either before the target was evaluated (Experiment 7) or before the target stimulus was even presented (Experiment 8). In this way influence awareness was measured before an overt evaluation took place or a covert evaluation could even be formed. 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 short, our findings demonstrate 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. Thus, when it comes to the AMP, that which is useful (influence awareness) is not particularly implicit, and that which is implicit (non-influence awareness) is not particularly useful.

Implications

Implicit vs. Explicit Accounts

Our findings consistently show that the majority of variance in group-level AMP effects is explained by those trials wherein participants are aware of how the prime will (prospective measures) or has (retrospective measures) influence(d) their evaluations. This claim holds across eight preregistered studies, different attitude domains, multiple versions of the AMP (standard, Mann et al., IA-AMP), and different influence awareness measures (prospective and retrospective measures taken on each trial vs. post hoc self-report questions). Thus it appears that the AMP effect is not implicit in the way that has previously been claimed (unaware), thus contradicting previous claims in this regard. Rather our findings are more consistent with an explicit account which argues that people are aware of the prime's influence on how they are responding to the target. Note that we are agnostic to the AMP's implicitness in other senses of the word (e.g., intentional): our goal here was to reevaluate the AMP's specific claim to implicitness in the sense of unawareness.

Theoretical Implications: Do AMP Effects Reflect A Misattribution Process?

So far we have focused on the 'implicitness' of AMP effects in terms of awareness. However, our findings are also relevant to another issue, namely the idea that AMP effects are mediated at the mental level by misattribution of prime valence to the target stimulus. Misattribution is traditionally conceived of as occurring in the absence of awareness (Schwarz & Clore, 1983; Payne et al., 2005). Indeed, as one reviewer of this manuscript noted, misattribution by definition cannot occur with awareness. If AMP effects rely heavily on awareness of prime influence (as our results indicate), then this suggests two possibilities.

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. However, such an approach runs contrary to how misattribution is traditionally defined (Schwarz & Clore, 1983), and would require a radical overhaul of the concept itself. Even if a redefinition of the construct were undertaken, 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.

On the other hand, it may be that misattribution is not the mechanism which mediates AMP effects. This possibility would have significant implications for a variety of theories and methods that rest on this idea. For instance, it would seriously challenge the misattribution account of AMP effects. It would call into question recent theoretical perspectives on misattribution that rely on the AMP for support. This includes theoretical models relating to the process of misattribution itself (e.g., the process model of misattribution: Payne, Hall, Cameron, & Bishara's, 2010), as well as claims that evaluative conditioning is based on a misattribution process (Jones et al., 2009), and that psychological properties beyond evaluations can also be misattributed (Blaison, Imhoff, Hühnel, Hess, & Banse, 2012). It would also call into question a number of second-generational tasks that attempt to exploit the misattribution of meaning (the Semantic Misattribution Procedure: Sava et al., 2012) and truth (the Truth Misattribution Procedure: Cummins & De Houwer, 2019). It seems likely that the very same issues associated with influence awareness in the standard AMP are likely to play similar roles in these other procedures. Future work could employ a similar IA-AMP style manipulation to these variants to investigate this issue in more detail. In short, our findings call into question the misattribution mechanism assumed to underpin AMP effects.

Practical Implications: Is The AMP A Valid Measure Of Attitudes?

Imagine that we set the AMP's status as an *implicit* measure to one side and merely ask the question: does the task have utility as a measure of attitudes in general? Our findings suggest it does not. One of the most pressing issues raised by our experiments is that instead of capturing general processes taking place in the general population, AMP effects seem to measure a subset of influence aware trials, especially in highly influence aware participants who are consistent across AMPs. In other words, 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). Such a finding suggests that scores on the measure do not reflect what most researchers assume or desire. This is highly problematic for its use in both basic and applied settings.

To illustrate, imagine that a researcher wants to assess implicit racial bias in law enforcement officers. She administers a race AMP to police officers, finds evidence of a large AMP effect at the group level, and subsequently infers that police officers are, in general, implicitly racially biased. Our findings suggest that such an AMP would not capture racial bias *in general*, but rather reflect the performance of a subset of participants who are highly aware that race-related primes were influencing their responses to the target stimuli. Importantly, these participants are likely to demonstrate AMP effects regardless of the domain being assessed. This is neither what is likely to be inferred from such a study nor what the researchers set out to capture. Put simply, most researchers who employ the AMP are interested in a given population's (implicit) evaluations and this is not what the task appears to measure.

Of course, one might contend that this issue applies to other implicit measures as well: effects in measures like the IAT, for example, could also be produced by a small subset of

consistent individuals. We agree. Although this study is (to our knowledge) the first of its kind to systematically investigate the correlation between implicit measure scores of the same individuals across multiple different domains, other measures could very well also exhibit similar issues. Investigating the presence and propensity of this issue for measures other than the AMP represents an important objective which should be pursued in future work. This would also help to better contextualize the AMP's susceptibility to this issue compared to other measures.

Just as our findings suggest that the presence of an AMP effect at the group level is not reflective of a general process in the general population, they also suggest that the AMP effect of a given individual (or lack thereof) is not diagnostic of that individual's evaluations. To illustrate, consider our previous example of implicit racial bias in police officers. Upon administering a race AMP to a specific police officer, the researcher observes that this specific officer displays a neutral AMP effect (i.e., they evaluate targets as equally pleasant when preceded by a Black face or a White face). The researcher concludes that this officer is less biased against black people compared to his contemporaries who, on average, demonstrate moderate anti-Black AMP effects.

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. If so, then the researcher's conclusions may be inappropriate. In short, our findings suggest that the absence of an AMP effect cannot be used to infer the absence of evaluations, which raises questions about the validity of the AMP itself.

Future Research Directions

Creating a Better Implicit Measure

One option is to modify the AMP effect in ways that exclude influence aware trials or refine the task itself in some way that diminishes the role of influence awareness on those effects. These changes would allow the AMP to maintain its status as an implicit measure. Our results could be seen as supporting this approach given that (a) even those participants with influence awareness rates of zero demonstrated (very small) IA-AMP effects, and (b) IA-AMP effects calculated from non-influence aware trials still possessed some predictive validity for discriminating between known groups. As such, researchers may be tempted to set the standard AMP to the side, employ an IA-AMP, exclude all influence aware trials, and calculate an effect. This is certainly one way forward. Yet it also comes with issues.

First, one should not conflate '*non*-influence aware' with 'influence *unaware'* responding. The IA-AMPs used here 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.

Yet even an IU-AMP would not be without issue. Imagine that an applied researcher in a specific domain wishes to examine differences between two known-groups using the IU-AMP and obtains results similar to what we report in Experiment 5 (d = 0.62, using IA-AMP non-influence aware trials for comparison). To appropriately power her study using the IU-AMP to

detect group differences would require at least 138 participants.[9] For the applied researcher, collecting such sample sizes is often either unfeasible or a poor use of limited resources. In contrast, if predictive utility was more important to her than 'process purity', then an IA-AMP capturing influence aware responses could detect group differences with as few as 16 participants (i.e., IA-AMP effects calculated from influence aware trials: $d = 2.08$).

Now imagine the flipside. For basic researchers, the need to collect larger sample sizes may be both feasible and desirable if this allows them to study implicit processes in a relatively 'pure' manner. The problem here is that an IU-AMP will likely also lead to a significant number of people being discarded due to zero, or near-zero, levels of unaware task responding. The implication here is that although such an IU-AMP might provide a better implicit measure by implementing changes to the task, the effects obtained from such a task would still not reflect behaviors (or mental processes) in people *in general*. Yet this is exactly what the AMP is primarily used for. Therefore, just as other fields acknowledge the variety of issues associated with making inferences or generalizations about people in general from non-representative samples (e.g., WEIRD individuals: Henrich, Heine, & Norenzayan, 2010; neuroscience tending to only study the brains of right-handed people: Willems, der Haegen, Fisher, & Francks, 2014; animal models of pathology that are based on male biology but not female: Mogil, 2016), we need to do the same. Both applied and basic researchers using the AMP (or AMP-like tasks, including the IU-AMP) need to carefully attend to the dangers of making inferences and generalizations about people *in general* from an effect that reflects a special subset of people.

Revision of Existing Findings

[9] Using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009): Independent *t*-test, alpha = 0.05 (two-sided), power = 0.95.

Assuming we are correct, our findings suggest that past conclusions made in the AMP literature may need to be revised. These conclusions are typically made on the basis of two common assumptions: (a) that AMP effects are reflective of implicit attitudes, and (b) that AMP effects represent an equally valid measure of such attitudes across all individuals (e.g., Fox et al., 2018; Kalmoe & Piston, 2013; Mann et al., 2019; Payne et al., 2005; Rinck & Becker, 2007; Spring & Bulik, 2014). To illustrate, consider a study by Franklin, Puzia, Lee, and Prinstein (2014), which concluded that "young adults with a history of non-suicidal self-injury (NSSI) display a significantly stronger *implicit* identification with [images of skin] cutting" compared to their counterparts without such a history of NSSI. Our findings suggest that such a result should instead be interpreted as "in those young adults who are highly influence aware on the AMP, those with a history of NSSI also self-identify more with NSSI compared to those who had no such history. However, little can be said about those with low influence awareness rates." This is just one example; similar revisions need to be applied to the core claims of all published research using the AMP (e.g., via systematic review), which may fundamentally alter the conclusions derived from that body of work.

What Makes A Person Influence Aware?

Our results also raise the question of what characterizes and differentiates highly influence aware participants who moderate AMP effects from the rest of the population. Experiment 6, which employed the modifications to the AMP suggested by Mann et al. (2019), suggests that the both influence awareness rate and moderation of the magnitude of AMP effects by influence awareness is not reduced through simple alterations to the task itself. Experiment 5 suggests that an individual's influence awareness rate is consistent across IA-AMPs assessing different domains, and that the influence awareness rate demonstrated in one domain predicts

AMP effects in another. As such, it seems that influence awareness rates are an individual difference variable rather than merely random noise or properties of the task itself. While beyond the scope of the current research, future work could examine whether influence awareness is a state- or trait-like property (e.g., whether it is consistent across time and context), whether it is related to other individual differences (e.g., Need for Cognition: Cacioppo & Petty, 1982), or indeed whether influence awareness on the AMP is related to performance on other kinds of implicit measures (e.g., the Implicit Association Test).

Diversity and Inclusiveness

Experiments 1-8 were carried out online on a platform that recruits participants from the general population (Prolific Academic). An analysis of the demographic data we requested in our studies (age, gender, and political orientation [in Experiments 4 and 5]) revealed that our samples were broadly representative in terms of age (ranging from 18 to 65 years), and balanced in terms of gender (847 women, 741 men, 15 other). As such, in our efforts to reexamine different properties of the AMP effect, we recruited samples that were, at worst, no less representative as the original AMP studies that we build and extend upon. At best, our samples are likely more representative than the original studies given that we did not recruit solely from undergraduate students (i.e., more balanced in terms of age and levels of education). That said, we did not request information on other demographic variables (e.g., sexuality, ethnicity). Although we had no theoretical reason to assume that such variables would moderate performance on a generic valence AMP, it is perhaps plausible they are associated with performance on the political IA-AMPs in particular. Likewise, given that Experiments 4-5 were sampled exclusively from US residents, and that the majority of Prolific Academic participants reside in the UK, our samples are primarily made up of (and potentially over-represent) individuals from these nations. Future

work may therefore wish to capture more detailed demographics information or could replicate our findings across different nationalities to further expand its remit.

Conclusion

AMP effects are not implicit in at least one way that they have previously been argued to be (i.e., unawareness). Our results show that both the magnitude of AMP effects and their predictive validity are strongly moderated by awareness. As such, insofar as (1) the AMP has been claimed to be implicit in the sense of being unaware and (2) utility can be defined in terms of large AMP effects and/or AMP effects that possess predictive validity (i.e., the criteria employed in this previous literature to date; e.g., Payne et al., 2013), what is useful about the effect is not particularly implicit, and what is implicit about the effect is not particularly useful. This finding raises a host of conceptual, theoretical, and applied issues for past and future research using the task as well as its ability to make inferences about psychological phenomena in people *in general* (i.e., rather than merely in a subset of highly aware participants).

References

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality & Social Psychology Bulletin*, *38*, 1194–1208. https://doi.org/10.1177/0146167212446835

Bar-Anan, Y., & Nosek, B. A. (2016). Misattribution of Claims: Comment on Payne et al., 2013. https://doi.org/10.31234/osf.io/r75xb

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Blaison, C., Imhoff, R., Hühnel, I., Hess, U., & Banse, R. (2012). The Affect Misattribution Procedure: hot or not? *Emotion*, *12*, 403–412. https://doi.org/10.1037/a0026907

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science, 10*, e1501.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Chapman, M. V., Hall, W. J., Lee, K., Colby, R., Coyne-Beasley, T., Day, S., … Payne, B. K. (2018). Making a difference in medical trainees' attitudes toward Latino patients: A pilot study of an intervention to modify implicit and explicit attitudes. *Social Science & Medicine, 199*, 202–208. https://doi.org/10.1016/j.socscimed.2017.05.013

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review, 24*, 212-232.

Cummins, J., & De Houwer, J. (2019). An inkblot for beliefs: The truth misattribution procedure. *PloS one, 14*, e0218661.

De Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers & A. W. Stacy (Eds.), *The Handbook of Implicit Cognition and Addiction* (pp. 11–28). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412976237.n2

Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology*, *13*, 120–126. https://doi.org/10.20982/tqmp.13.2.p120

Ditonto, T. M., Lau, R. R., & Sears, D. O. (2013). AMPing Racial Attitudes: Comparing the Power of Explicit and Implicit Racism Measures in 2008. *Political Psychology, 34*, 487–510. https://doi.org/10.1111/pops.12013

Dunham, Y., & Emory, J. (2014). Of affect and ambiguity: The emergence of preference for arbitrary ingroups. *Journal of Social Issues, 70*, 81–98. https://doi.org/10.1111/josi.12048

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

Fox, K. R., Ribeiro, J. D., Kleiman, E. M., Hooley, J. M., Nock, M. K., & Franklin, J. C. (2018). Affect toward the self and self-injury stimuli as potential risk factors for nonsuicidal self-injury. *Psychiatry Research, 260*, 279–285. https://doi.org/10.1016/j.psychres.2017.11.083

Franklin, J. C., Puzia, M. E., Lee, K. M., & Prinstein, M. J. (2014). Low Implicit and Explicit Aversion Toward Self-Cutting Stimuli Longitudinally Predict Nonsuicidal Self-Injury. *Journal of Abnormal Psychology, 123*, 463–469.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd, *Handbook of Research Methods in Social and Personality Psychology* (pp. 283–310). Cambridge University Press.

Gawronski, B., & Payne, B. K. (2010). *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. Guilford Press.

Gawronski, B., & Ye, Y. (2014). What Drives Priming Effects in the Affect Misattribution Procedure? *Personality and Social Psychology Bulletin, 40*, 3–15. https://doi.org/10.1177/0146167213502548

Görgen, S. M., Joormann, J., Hiller, W., & Witthöft, M. (2015). The Role of Mental Imagery in Depression: Negative Mental Imagery Induces Strong Implicit and Explicit Affect in Depression. *Frontiers in Psychiatry*, *6*, 94. https://doi.org/10.3389/fpsyt.2015.00094

Greenwald, A., McGhee, D., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Hermans, D., De Houwer, J., & Eelen, P. (1994). The affective priming effect: Automatic activation of evaluative information in memory. *Cognition and Emotion*, *8*(6), 513–533. https://doi.org/10.1080/02699939408408957

Imhoff, R., Schmidt, A. F., Bernhardt, J., Dierksmeier, A., & Banse, R. (2011). An inkblot for sexual preference: A semantic variant of the Affect Misattribution Procedure. *Cognition and Emotion*, *25*(4), 676–690. https://doi.org/10.1080/02699931.2010.508260

Jasper, F., & Witthöft, M. (2013). Automatic Evaluative Processes in Health Anxiety and Their Relations to Emotion Regulation. *Cognitive Therapy and Research*, *37*(3), 521–533. http://dx.doi.org/10.1007/s10608-012-9484-1

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of Personality and Social Psychology*, *96*(5), 933–948. https://doi.org/10.1037/a0014747

Kalmoe, N. P., & Piston, S. (2013). Is Implicit Prejudice against Blacks Politically Consequential? Evidence from the AMP. *Public Opinion Quarterly*, *77*(1), 305–322. https://doi.org/10.1093/poq/nfs051

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. NIMH Center for the Study of Emotion and Attention.

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology*, *116*(3), 349–374. https://doi.org/10.1037/pspa0000146

Mann, T. C., & Ferguson, M. J. (2017). Reversing Implicit First Impressions through Reinterpretation after a Two-Day Delay. *Journal of Experimental Social Psychology*, *68*, 122–127. https://doi.org/10.1016/j.jesp.2016.06.004

McCarthy, R. J., Skowronski, J. J., Crouch, J. L., & Milner, J. S. (2017). Parents' spontaneous evaluations of children and symbolic harmful behaviors toward their child. *Child Abuse & Neglect*, *67*, 419–428. https://doi.org/10.1016/j.chiabu.2017.02.005

Mogil, J. S. (2016). Perspective: Equality need not be painful. *Nature*, *535*, S7

Moors, A., & De Houwer, J. (2006). *Automaticity: A theoretical and conceptual analysis*. *Psychological Bulletin*, *132*, 297–326. https://doi.org10.1037/0033-2909.132.2.297

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, D., & Lundberg, K. (2013). Intention Invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, *39*(3), 375–386. https://doi.org/10.1177/0146167212475225

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., & Lundberg, K. (2014). The Affect Misattribution Procedure: Ten Years of Evidence on Reliability, Validity, and Mechanisms. *Social and Personality Psychology Compass*, *8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, *51*(2), 109-127. DOI: https://doi.org/10.5334/pb-51-2-109

Rinck, M., & Becker, E. (2007). Approach and avoidance in fear of spider. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(2), 105–120. https://doi.org/10.1016/j.jbtep.2006.10.001

Sava, F. A., MaricutToiu, L. P., Rusu, S., Macsinga, I., Vîrgă, D., Cheng, C. M., & Payne, B. K. (2012). An inkblot for the implicit assessment of personality: The semantic misattribution procedure. *European Journal of Personality*, *26*(6), 613–628. https://doi.org/10.1002/per.1861

Schreiber, F., Witthöft, M., Neng, J. M. B., & Weck, F. (2016). Changes in negative implicit evaluations in patients of hypochondriasis after treatment with cognitive therapy or exposure therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, *50*, 139–146. https://doi.org/10.1016/j.jbtep.2015.07.005

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgements of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523.

Smith, A. R., Forrest, L. N., Velkoff, E. A., Ribeiro, J. D., & Franklin, J. (2018). Implicit attitudes toward eating stimuli differentiate eating disorder and non-eating disorder groups and predict eating disorder behaviors. *The International Journal of Eating Disorders*, *51*(4), 343–351. https://doi.org/10.1002/eat.22843

Spring, V. L., & Bulik, C. M. (2014). Implicit and explicit affect toward food and weight stimuli in anorexia nervosa. *Eating Behaviors*, *15*(1), 91–94. https://doi.org/10.1016/j.eatbeh.2013.10.017

Teige-Mocigemba, S., Becker, M., Sherman, J., Reichardt, R., & Klauer, K. C. (2017). The Affect Misattribution Procedure: In Search of Prejudice Effects. *Experimental Psychology*, *64*(3), 215–230. https://doi.org/10.1027/1618-3169/a000364

Van Dessel, P., Mertens, G., Smith, C. T., & De Houwer, J. (2017). The Mere Exposure Instruction Effect. *Experimental Psychology*, *64*(5), 299–314. https://doi.org/10.1027/1618-3169/a000376

Ye, Y., & Gawronski, B. (2018). Validating the Semantic Misattribution Procedure as an Implicit Measure of Gender Stereotyping. *European Journal of Social Psychology*, *48*(3), 348–364. https://doi.org/10.1002/ejsp.2337

Zerhouni, O., Bègue, L., Comiran, F., & Wiers, R. W. (2018). Controlled and implicit processes in evaluative conditioning on implicit and explicit attitudes toward alcohol and intentions to drink. *Addictive Behaviors*, *76*, 335–342. https://doi.org/10.1016/j.addbeh.2017.08.026