British Journal of Social Psychology (2017)

The British Psychological Society

www.wileyonlinelibrary.com

Stress-testing the affect misattribution procedure: Heterogeneous control of affect misattribution procedure effects under incentives

Chad J. Hazlett[1]* and Adam J. Berinsky[2]

[1]Department of Political Science, Department of Statistics, University of California, Los Angeles, California, USA

[2]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

The affect misattribution procedure (AMP) is widely used to measure sensitive attitudes towards classes of stimuli, by estimating the effect that affectively charged prime images have on subsequent judgements of neutral target images. We test its resistance to efforts to conceal one's attitudes, by replicating the standard AMP design while offering small incentives to conceal attitudes towards the prime images. We find that although the average AMP effect remains positive, it decreases significantly in magnitude. Moreover, this reduction in the mean AMP effect under incentives masks large heterogeneity: one subset of individuals continues to experience the 'full' AMP effect, while another reduces their effect to approximately zero. The AMP thus appears to be resistant to efforts to conceal one's attitudes for some individuals but is highly controllable for others. We further find that those individuals with high self-reported effort to avoid the influence of the prime are more often able to eliminate their AMP effect. We conclude by discussing possible mechanisms.

The affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) is a powerful technique used to measure preferences and attitudes in a wide variety of contexts where people may wish to conceal them (see Gawronski & Payne, 2010 for review). These include racial bias (Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009; Kalmoe & Piston, 2013), emotional responses to smoking cues (Payne, McClernon, & Dobbins, 2007), self-esteem (Schreiber, Bohn, Aderka, Stangier, & Steil, 2012), and hedonic responses to food (Hofmann, van Koningsbruggen, Stroebe, Ramanathan, & Aarts, 2010) and alcohol (Payne, Govorun, & Arbuckle, 2008). The AMP is a priming-based paradigm in which visual primes evoke affect, influencing participants' appraisal of the pleasantness of subsequent neutral target images. Comparing the proportion of positive target appraisals for different categories of primes reveals differences in positive or negative affect activated by each prime category.

The AMP has several attractive features, including the large magnitude of the priming effect and its high statistical reliability (see e.g., Payne *et al.*, 2005; which produces Cronbach-α measures in the range of in .85–.95). Relative to other widely used measures thought to reveal attitudes people seek to conceal, such as the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), the AMP is relatively easy for researchers to

*Correspondence should be addressed to Chad J. Hazlett, Department of Political Science, Department of Statistics, University of California, L3264 Bunche Hall, Los Angeles 90095, CA, USA (email: chazlett@ucla.edu).

DOI:10.1111/bjso.12217

implement and for participants to complete. For a review of these benefits, see Gawronski and De Houwer (2014).

Recent studies have examined the mechanism underlying the AMP effect, such as whether it reflects affective or semantic processing of the prime, and whether it reflects implicit attitudes of which the participant is unaware (e.g., Bar-Anan & Nosek, 2012; Gawronski & Ye, 2014; Payne, Hall, Cameron, & Bishara, 2010; Payne *et al.*, 2013). However, regardless of the mechanism at work, investigators typically use the AMP or the IAT not to isolate implicit from explicit attitudes or other theoretical considerations, but simply because they are asking questions where self-presentational concerns may influence the results.

The AMP is a promising tool for such tasks, because it is thought to 'measure influences of attitudes on behavior that persist in opposition to participants' intentions' (Payne *et al.*, 2005; p 278). This purported inability of participants to control their responses, despite awareness of their attitudes, is theorized to stem from a misattribution mechanism: participants are simply unable to disentangle their reaction to the target from that of prime (Payne, Hall, *et al.*, 2010; see Gawronski & Ye, 2014 for review).

While the usefulness of the AMP as a measure of sensitive attitudes relies critically on resistance to manipulation, the empirical evidence for this resistance remains underdeveloped. Until recently, the primary evidence stemmed from initial work by Payne *et al.* (2005), finding that (1) the AMP effect is not (significantly) diminished by warning participants about the effect of the prime on their assessments of the target, and (2) AMP effects reveal racial prejudices that most individuals would prefer to conceal if possible. We argue that these findings provide suggestive but incomplete evidence for the AMP's resistance to corrective efforts. Most critically, the warnings used in these experiments may not have been sufficient to motivate participants to conceal their reactions to the prime. As Payne *et al.* (2005) themselves note, the warning might have failed because, 'despite being aware of the potential for bias, participants were not motivated enough to change their behavior' (281). This concern is particularly relevant given that the AMP is most needed as a measure of sensitive attitudes precisely when participants have powerful incentives to manipulate their responses.

Other work has called the AMP's resistance to control into question. Teige-Mocigemba, Penzl, Becker, Henn, and Klauer (2015) show that in an experiment with a cover story encouraging participants to evaluate prime images of Arabs as positive and celebrities as negative – opposite to the expected biases – participants were able to produce AMP effects in the instructed direction. This suggests that AMP results can be controlled, particularly by effortfully altering one's evaluation of the prime. However, this study leaves important questions unanswered. While Teige-Mocigemba *et al.* show that participants can manipulate their AMP effects, it does so by instructing subjects to evaluate (or appear to evaluate) the primes in certain ways, opposite to the stereotypical expectation. The study likely encouraged participants to increase their emphasis on evaluating the prime in the instructed direction, constituting a particular strategy for influencing the AMP response. It thus remains an open question as to whether individuals can manipulate their response to the targets to *disagree* with their evaluation of the prime, as we would expect participants to attempt in a standard implementation of the AMP.

To this end, Eder and Deutsch (2015) sought to determine whether increasing participants' motivation to evaluate the target and not the prime could reduce or eliminate the AMP effect. They use target images that appear to have a 'correct' interpretation as positive or negative and give feedback on whether participants correctly classified the

targets. In doing so, they find a significant AMP effect remains, though diminished in size, suggesting that 'participants have at least partial control over a priming influence'.

Our work is most similar in spirit to Eder and Deutsch (2015), in that we seek to determine whether participants can avoid the biasing effect of primes when motivated to do so. However, we sought to test this with as few changes to the AMP as possible, to determine whether, in a standard AMP procedure used to measure attitudes, participants are able to avoid revealing their attitudes towards the prime through effortful control, without a change in cover story or instructions.

Materials and methods

The standard warning instructions for the AMP (following Payne *et al.*, 2005) include the warning that, 'Sometimes, the [prime image] presented prior to the [target image] can bias your responses on [the target image]. Thus, please try to make sure that your responses are not influenced by the [prime image]'. Our supposition is that being asked to 'try to make sure' the prime has no effect is insufficient motivation for most participants. Such a warning instruction may not provide a level of motivation akin to an individual's motivation to conceal a socially undesirable attitude, for example. Accordingly, finding that the AMP effect persists under this instruction is not strong evidence for its resistance to effortful control. While Eder and Deutsch sought to increase motivation to control by providing feedback on 'correct' or 'incorrect' classifications, we used the same primes, targets, and instructions as in Payne *et al.* (2005). We only added to the instructions a monetary incentive for participants to rate the targets 'honestly' as pleasant or unpleasant, without suggesting any particular response strategy to the participants.

Our implementation of the AMP closely follows experiments 1–4 of Payne *et al.* (2005). We use as primes the images collected by Payne, which were selected to have near universal positive or negative appraisal. There is no obvious reason for individuals to conceal their affect towards these primes. Each trial begins with an affectively positive or negative prime, presented for a duration of 75 ms. After an interstimulus interval of 125 ms, the target (a Chinese character) is then presented for 200 ms, followed by a white noise mask. All stimuli are the same size and presented at fixation. The duration of the prime image (75 ms) and the interstimulus interval (125 ms) are those used throughout (Payne *et al.*, 2005). Our target duration of 200 ms was chosen after pilot testing of our online platform, and falls in between the 100 ms used in Payne *et al.* (2005) and the longer 250-ms duration used in an online version of the AMP (Payne, Krosnick, *et al.*, 2010).

Participants are instructed to press the 'E' key on the left if they judge the Chinese character to be pleasant or the 'I' key on the right if they judge it to be unpleasant. Each participant undergoes 30 trials, using images selected randomly with replacement for both the prime and target.

Participants were recruited online through the Amazon Mechanical Turk (MTurk) platform. Paolacci and Chandler (2014) find that data collected from such samples are reliable (see also Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010).

After providing demographic information, participants followed a link to a custom online platform to complete the AMP task. Upon arriving there, participants were randomized into one of six conditions, each differing only in the instructions to the participant. To briefly describe each condition.

Basic

Participants are given the traditional AMP instructions, following those of Payne *et al.* (2005), without warning of the potential priming effect. The core part of these instructions states:

> 'This study examines how people make simple judgments. You will see pairs of pictures flashed one after another. The first is a real-life image. The second is a drawing. Your job is to judge the visual pleasantness of the drawing'.

Warning

Participants are warned against the potential priming effect, following the stronger warning treatment employed in experiment 2 of Payne *et al.* (2005):

> 'It is important to note that having just seen a positive image can sometimes make you judge the drawing more positively than you otherwise would. Likewise, having just seen a negative image can make you judge the drawing more negatively. . . Please try your absolute best not to let the real-life images bias your judgment of the drawings! Give us an honest assessment of the drawings, regardless of the images that precede them'.

Incentive

In four different incentivized conditions, participants were not only told to give an honest assessment but were incentivized by a bonus to avoid being influenced by the prime in their assessment of the target's pleasantness. The incentive was randomly chosen to be \$0.10, \$0.25, \$0.75, or \$1.00.

> 'To ensure you work hard to avoid being influenced by the images, you will be paid according to how honestly you rate the drawing as pleasant or unpleasant. We will analyze your data to determine how honestly you have rated the drawings. If we determine your answers are honest assessment and not influenced by the pictures, you will earn an extra [\$0.10/\$0.25/\$0.75/\$1.00]!'

The *Basic* and *Warning* instructions above are identical to those used in Payne *et al.* (2005); the *Incentive* instructions were chosen to be as similar as possible while clearly introducing the monetary incentive. While the incentives may seem small, respondents could significantly increase their pay over the baseline of \$0.40.

We emphasize that the *Incentive* language is only a more emphatic version of the existing *Warning* language previously used. As the *Warning* language asks for an 'honest assessment of the drawings, regardless of the images that precede them', the *Incentive* instruction repeats that language then adds an incentive, continuing to use the notion of an 'honest' response as one that is not influenced by the prime. We leave as implicit that the investigator will be able to make an assessment of a participant's honesty. In actuality, all participants in the *Incentive* arm are rewarded with the bonus.

Sample size and rejection criteria

Pilot experiments showed that even in small samples (40–50 participants per condition), differences between the *Incentive* and the other conditions began to reach significance. However, to maximize our ability to characterize these differences, we used the largest sample allowed by our budget, requesting 1,350 participants on MTurk. Out of this, we

received 1311 participants. We remove participants who completed fewer than 10 trials. To remove trials on which participants were clearly not paying attention, we reject trials on which reaction time exceeded 5 s, removing 4% of trials and 15 participants. We also remove participants who gave the same response to all targets (5%), and those who report being able to read or write in Chinese (5%).[1] This left 1144 subjects in analyses below: 182 in *Basic*, 189 in *Warn*, and 773 in the *Incentive* conditions (190 at \$0.10, 193 at \$0.25, 211 at \$0.75, and 179 at \$1.00). The effective completion rate once all these criteria are applied is very similar across conditions, with 90% in the *Basic* condition, 86% in the *Warning* condition, and 87% in the *Incentivized* conditions (85–90% within each incentive level). Because these completion rates are high, the differences between them are small, and we conclude that the following results are not due to differential attrition between the conditions.

Additional information

Following the AMP, participants completed several additional items. Demographic information included year of birth, gender, education, and whether the participant can read Chinese (for exclusion criteria). We also collected the two-item need for cognition scale (Cacioppo & Petty, 1982) and the two-item need for evaluation (Jarvis & Petty, 1996) scale. Finally, we collected a self-reported measure of the effort expended in attempting to respond to the target and not be influenced by the prime, described at length below. The Appendix shows the complete instrument.

Results

Average AMP effects by condition

We first describe the average size of the AMP effect by condition. Our analyses begin by computing individual-level AMP effects, by taking the proportion of positive responses when shown a positive prime, minus the proportion of positive responses following a negative prime. Figure 1 shows the mean AMP effect, by condition, with 95% confidence intervals.

As expected, the AMP effect is largest in the *Basic* condition, where no warning is given regarding the possibility of misattribution. Under this condition, participants are on average 35 [CI: 29, 41] percentage points more likely to rate a target as positive following a positive prime than following a negative prime. The effect remains large at 31 [CI: 28, 33] percentage points in the *Warning* condition.

To compare the mean effect size under each condition statistically, we simply regress the individual AMP effect estimates on an intercept and indicators for the *Warning* and *Incentive* conditions allowing heteroscedasticity robust ('HC1') standard errors. We note that this is numerically equivalent to conducting a *t*-test with unequal variances.

As in Payne *et al.* (2005), the means for *Warning* and *Basic* were not statistically distinguishable ($t = 0.93$, $df = 1{,}141$, $p = .35$).

All four *Incentive* conditions had similar means and were statistically indistinguishable from each other. We therefore pool them into a single condition. The mean AMP effect under *Incentive* is 22 [CI: 17, 28] percentage points. This mean is significantly lower than

[1] *As long as participants giving the same response on every trial are removed, all analyses can be replicated on the original data without the other rejection criteria used here and produce very similar results with the same substantive conclusions.*

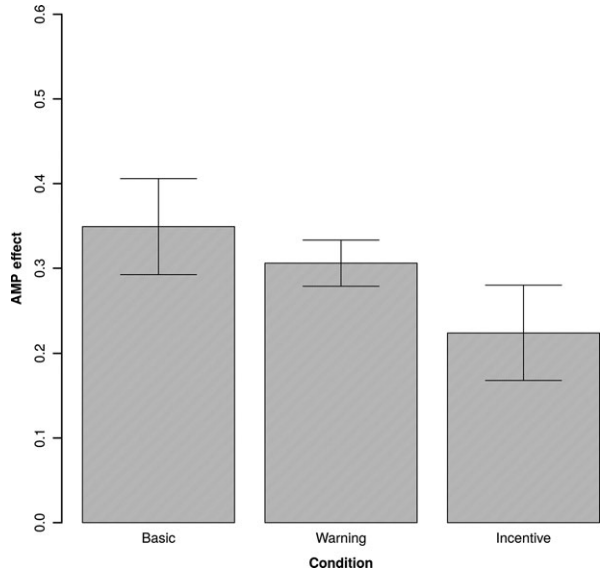

Figure 1. Mean AMP Effect Estimates, by Condition. *Note.* Average subject-level AMP effect by condition, after applying rejection criteria described in the text. Error bars indicate 95% confidence intervals. The mean AMP effect under the incentive condition is significantly lower than either of the other two conditions, despite overlap in the confidence intervals.

that in the *Basic* condition ($t = 3.59$, $df = 1{,}141$, $p < .001$) or in the *Warning* condition ($t = 1.98$, $df = 1{,}141$, $p = .047$).

In addition, while our main outcome of interest is the AMP effect, we also consider the effect of condition on reaction time (RT).[2] One may expect that under the *Incentive* and perhaps *Warning* conditions, participants spend extra time in an effort to 'undo' the effect of the prime image or refocus their attention on the target image. Simply regressing each participant's mean reaction time on indicators for the conditions, we find RTs in the *Basic* condition average 862 ms ($SE = 34$ ms), while RTs in the *Warning* condition are longer by 75 ms ($SE = 47$ ms), although the difference is insignificant ($t = 1.61$, $df = 1{,}141$, $p = .11$). Under the *Incentive* condition, the average RTs are 200 ms ($SE = 37$ ms) longer than in the *Basic* condition, a highly significant difference ($t = 5.61$, $df = 1{,}141$, $p < 10^{-6}$). This suggests that participants in the *Incentive* condition are apparently making a greater effort to follow the instructions of being influenced by the target rather than the prime image, at the expense of considerably longer RTs. Note that consistent with evidence that the *amount* of the incentive does not influence the AMP effect, within the *Incentive* condition, the amount of the incentive does not explain a significant component of the variation in RT ($F = 1.57$, $df1 = 3$, $df2 = 769$, $p = .20$).

Distributional effects

The above analyses used individual-level AMP effects, but considered only the mean of these effects by condition. In what follows, we look to the entire empirical distribution of

[2]*We thank an anonymous reviewer for suggesting this analysis.*

individual-level AMP effects under each condition, revealing considerably more information about how individuals respond to the *Basic*, *Warning*, and *Incentive* instructions.

Figure 2 shows the distribution of individual-level AMP effect estimates, by condition. Under the *basic* condition (long dash), the modal AMP effect is large, peaking at an effect size of approximately 50 percentage points. Note that this modal effect is higher than the mean effect of 37 percentage points for this group. This is because a smaller, lower mode appears at an effect size of approximately 0 percentage points, indicating a subpopulation for whom there is no evident AMP effect. Under the *Warning* condition (short dash), there is again one mode around an effect size of 40–50 percentage points. However, another significant portion of the observations fall below this mode, with a large contribution from a group centred around 0 percentage points. Thus, while the mean AMP effects under the *Basic* and *Warn* conditions were not distinguishable, the shape of the distribution for the two conditions do differ significantly (Kolmogorov–Smirnov two-sample $D = .166$, $p = .03$).

Finally, under the incentivized conditions (short dash), the lower mode near 0 percentage points now dominates, and is clearly separated from the higher mode, which remains at approximately 50 percentage points. Once again, the shape of the distribution of responses under *Incentive* is widely different from the distribution under *Warn* ($D = .17$, $p = .001$), or *Basic* ($D = .21$, $p < 10^{-4}$).

This analysis reveals that the incentives – and to a lesser degree the warnings – reduced the mean AMP effect shown above, but not simply by shifting the whole distribution of AMP effects towards zero. Rather, in all conditions, one subgroup responds strongly with an effect of roughly 50 percentage points, while another has an effect near 0 percentage points. Put differently, increasing the incentives to resist the AMP effect does not reduce the size of the effect among the group of respondents where it occurs, but rather increases the proportion of participants entirely concealing their response to the prime. This has important implications: not only is there some degree of 'beating the test', but the proportion of people who beat the test differs depending on condition.

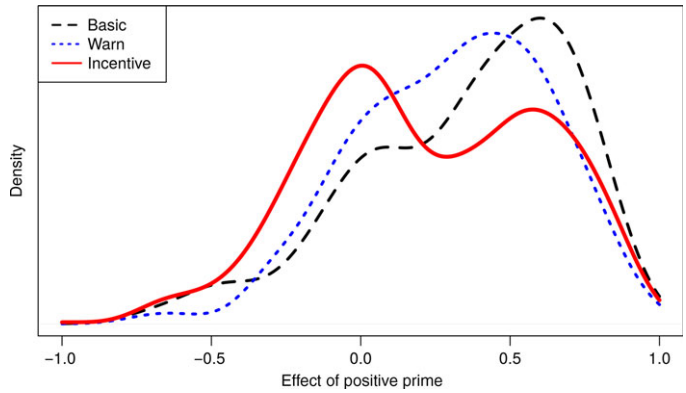

Figure 2. Distribution of individual-level AMP effects, by condition. *Note.* The 'basic' condition (long dash) shows a large mode near an effect size of 0.5, corresponding to a 50 percentage point effect), but also shows a small mode near an effect of zero. Moving to the 'warn' condition (short dash), fewer individuals are in the higher mode and a larger group appears in the mode over zero. Finally, in the 'incentive' condition, the mode over zero is now the larger mode, though a mode near the full effect size of 50 percentage points remains.

Individual characteristics: Self-reported effort

If some individuals are able to evade the priming effect on which the AMP is based while others are not – as our analysis finds – it would be useful to know which individuals fall into each of these groups. All of the individual-level traits measured during the experiment – education, age, gender, need for evaluation, and need for cognition – failed to predict individuals' AMP effects. However, at the end of the survey, we asked participants whether they tried 'extremely hard' (5), 'hard' (4), 'somewhat hard' (3), 'slightly hard' (2), or 'did not try at all' (1) to prevent the prime from influencing their judgement of the target character. It was made clear that payment and bonus decisions would not be affected by responses to this question.

The effort reported by each participant was strongly related to mean AMP effects: a one unit higher level of effort is associated with an 8.9 [CI: 4.8, 13] percentage point lower AMP effect in the *Warning* condition, and 8.0 [CI: 5.6, 10] percentage point reduction under the *Incentive* condition ($p < 10^{-5}$ for each, separately).[3] Relatedly, simply being at or above the median level of self-reported effort (a '4' on the scale) rather than below, it predicts a 17 percentage point smaller AMP effect in the *Warn* condition and a 19 percentage point smaller AMP effect in the *Incentive* condition ($p < .001$ for both) – very substantial changes.

Despite the size of these differences in mean AMP effects, the shift in means is again only part of the story, and self-reported effect (as a linear predictor) explains only approximately 6% of the variance in AMP effects. It is again more illuminating to compare the entire distribution of AMP effect sizes under different levels of self-reported effort. Splitting the sample in each condition at the median level of reported effort for that condition, we see that high-effort individuals are more often those with little or no AMP effect, while low-effort individuals are more often those with unmitigated AMP effects (Figure 3). This relationship is most apparent in the *Incentive* condition, as would be expected. However, the *Warning* condition shows a similar, weaker relationship. We note that self-reported effort was unrelated to need for cognition, need for evaluation, and education (all with $p > .20$), ruling out indirect effects of these characteristics on the AMP effect through self-reported effort.

Finally, participants' self-reported effort was also strongly related to their average reaction times, but only in the *Incentive* and *Warn* conditions. Participants reporting median or higher effort ('4' or '5') took 157 ms longer to respond in the *Warn* condition ($t = 2.42, p = .017$), and 128 ms longer in the *Incentive* condition ($t = 3.40, p < .001$). There is no significant relationship between reaction time and self-reported effort in the *Basic* condition. Our discussion of self-reported effort and its usefulness below is largely agnostic as to what it actually measures. However, we note that this relationship between self-reported effort and reaction time suggests that the former may indeed index the effort participants expend to avoid being influenced by the prime, as such effortful manipulation likely takes time.

Discussion

Our results first replicate key findings of Payne *et al.* (2005): we show that warning participants about the effect of the prime on subsequent judgements in the AMP does not eliminate or significantly mitigate the average AMP effect. The applicability of the AMP for the measurement of sensitive attitudes, however, relies on a stronger assumption that this

[3] *These effects easily survive Bonferroni correction for multiple comparisons.*

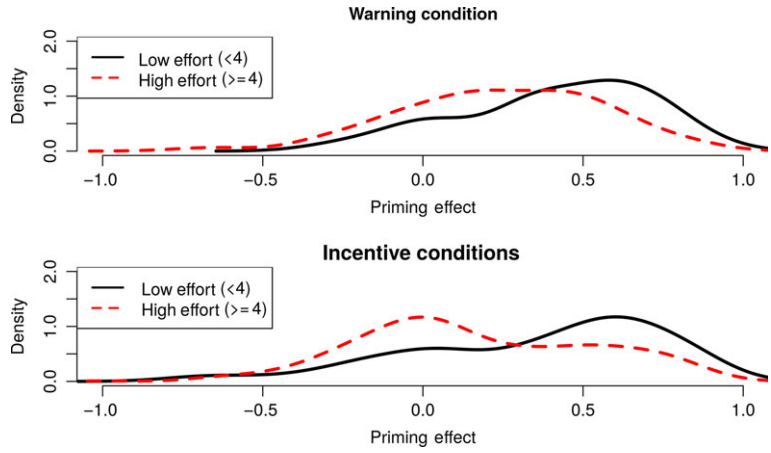

Figure 3. Distributions of Individual AMP Effects, by Effort. *Note.* Full distribution of AMP effects under both 'warning' and 'incentive' conditions, by median split on self-reported effort to avoid the influence of the prime on their assessments of the targets. Under the 'warning' condition, those reporting lower than median effort show the largest AMP effects, while those at higher than median effort have more individuals with lower AMP effects. The distinction becomes much clearer under the 'incentive' condition: those reporting lower effort are most concentrated at the full effect size, while those reporting higher effort show much smaller effects with the largest concentration around zero effect.

effect persists when participants are strongly motivated to suppress it. We therefore sought to determine whether participants could conceal their affect towards the primes when given a stronger incentive to do so. We find that participants do indeed have a lower average AMP effect under such incentives.

The findings are consistent with existing work. First, De Houwer and Smith (2013) find that instructing responds to rely on their spontaneous, 'gut' reactions to the target enhances the AMP effect size. One explanation is that this is effectively the opposite instruction to what we give, reducing efforts to respond only to the content of the target while filtering out the effect of the prime.[4]

Second, our results are consistent with those of results of Teige-Mocigemba *et al.* (2015) although, unlike us, they use a cover story to motivate participants to show attitudes towards the primes opposite of those typically expected. Our approach differs principally in that we do not ask participants to change their perception of the primes, but rather to disallow the primes from influencing their responses. This may be more in keeping with the strategies that participants might invoke to conceal sensitive attitudes.

Third, the result is also broadly consistent with Eder and Deutsch (2015), who show that motivating participants to evaluate the target rather than prime by providing feedback on allegedly 'correct' or 'incorrect' evaluations reduces the AMP effect. Our design differs in our direct replication of Payne *et al.* (2005), adding only incentives to follow the existing instructions. We do this to replicate the experience participants would have in an AMP experiment designed to measure sensitive attitudes, without a cover story or instructions that suggest a strategy for beating the test.

In a 'sensitive attitudes' application of the AMP, the motivation to control one's response would be supplied by self-presentation concerns. Here, in order to know the

[4]*We thank an anonymous reviewer for noting this connection.*

'ground truth' AMP effect, we use Payne's original primes with near universal positive or negative valence, instead motivating respondents to control their response through the monetary incentive. Our supposition is that if a very small monetary incentive (even $0.10) is sufficient to compel some participants to eliminate their AMP effect, participants' self-presentational concerns will likely be more than sufficient.

Although one might expect to see a greater reduction in the mean AMP effect as the incentive increases, over the range of incentives tested here ($0.10 to $1.00), we find that only the presence or absence of an incentive matters, and not its size. Of course, larger incentives might change behaviour in more extreme ways. However, at the very least, our results demonstrate that modest incentives can change the behaviour of the AMP respondents. One reason why the amount of the incentive may not matter is that the *Incentive* condition could influence behaviour through placing additional attention on the instructions, or implying that participants will be more closely evaluated. For purposes of our question, this is not an important distinction: whether it is a few words or a few cents doing the work, we find that a subset of participants clearly are able to overcome their AMP effects. Even in the *Warning* condition as previously used (e.g., Payne *et al.*, 2005), the distributional analysis we provide shows that some participants overcome the effect. The proportion is simply increased through the *Incentive* instruction.

Heterogeneous control

A focus on the distribution of the AMP effects is critical for understanding how individuals respond to that task. Our results provide the first evidence of heterogeneity in control over responses, and this heterogeneity is central to understanding the effect of incentives on behaviour. First, even in the *Basic* condition, there is evidence that while many subjects show a significant AMP effect, a smaller but distinct 'no effect' mode exists with nearly zero AMP effect. This evidence has not previously been reported. Second, the *Warning* condition increases somewhat the portion of the sample falling into that 'no effect' mode. While neither this study nor Payne *et al.* (2005) found a significant effect of the warning on the average AMP effect, we find that the *Warning* condition does significantly alter the distribution of outcomes. Third, the *Incentive* condition reduces the average AMP effect, again not by reducing the AMP effect uniformly across participants, but by placing a larger portion of the sample into the (existing) 'no effect' mode. This heterogeneity in the level of control – with some individuals obtaining full control – is particularly troubling: differences on AMP-derived measures between groups of people, for example, may reflect differences in controllability rather than in the attitudes the AMP is intended to measure.

While no individual traits we measured predict individuals' AMP effects, the level of effort individuals self-report in seeking to avoid influence by the prime is strongly predictive of the mode into which they fall. This is true across conditions. Whether this self-reported effort meaningfully reflects the effort participants spend – or is confabulated after the fact to explain their responses – remains an open area of debate (for a related discussion of self-reported intention to rate primes, see Bar-Anan & Nosek, 2012; Payne *et al.*, 2013; and Gawronski & Ye, 2014). For the narrow purposes of determining the

AMP's resistance to manipulation, we note simply that this self-reported effort item provides a means of predicting who is more likely to fall into the 'full effect' versus the 'no effect' mode under incentives to conceal responses to the prime.[5] It is also possible that there is variation in understanding of the instructions, such that some participants both failed to make efforts to avoid the influence of the prime and had large AMP effects.[6] That said, we see no relationship between education and AMP effect estimates or self-reported effort.

Our central concern was the 'fakeability' of the AMP, but our results allow some speculation into the mechanisms by which individuals control their response. First, we chose a design which makes the minimal changes to the standard AMP in order to increase effort participants spend in 'honestly' evaluating the target image. The simplest explanation for the reduced AMP effect given this change is that participants simply follow this instruction in our experiment because they have greater incentive to do so.[7] Prior work finding partial controllability of the AMP has not distinguished between an 'early modification' mechanism in which participants change how they evaluate the prime, and a 'late modification' mechanism in which they modify only their response. We do not seek to settle that debate here. However, we find suggestive evidence for the 'late modification' mechanism in this case. First, our design elicits control over the AMP without encouraging participants to alter their evaluations of the prime images, as in Teige-Mocigemba *et al.* (2015). It would seem surprising if participants choose to – or even could in this case – alter their immediate reactions to primes of such universal positive/negative valence. Thus, intentionally changing one's evaluation of the prime is at least not necessary to achieve control over the AMP response. Second, our finding that in the *Warn* and *Incentive* conditions, self-reported effort, reaction time, and control of the AMP effect are all correlated suggests that the process of avoiding the influence of the prime may involve both effort and time. Such a pattern would be surprising if the primes did not elicit an initial response that required later effortful control to overcome before responding.

More broadly, our results show that no elaborate instruction or cover story is required to coach participants into overcoming the AMP effect – simply offering an incentive to follow the instruction already built into the W*arning* condition (to not be influenced by the prime) is sufficient for some participants. We do not know what strategy participants use to achieve this control under these broad instructions. One strategy of particular concern is whether participants simply attempt to produce a seemingly random sequence of responses, unrelated to any perception of the primes or of the targets they are supposed to evaluate. This strategy was clearly not followed by subjects in Teige-Mocigemba *et al.* (2015), as those subjects were motived to and successful in reversing their AMP effects rather than zeroing them out. It also would not explain the results of Eder and Deutsch (2015), in which there were putatively 'correct' answers for the classification of the targets. Here, further analyses of our data find that few subjects show evidence of

[5] *This process of utilizing information from self-reported effort – or other variables found to predict controllability – to characterize responses among those who are more or less able to control their responses could be formalized by a method such as latent class analysis.*

[6] *We thank an anonymous reviewer for this suggestion.*

[7] *Whether this increased ability to evaluate the target 'honestly' is achieved through increased attention to the target or by conscious efforts to 'de-bias' their affect by 'adjusting for' the prime remains uncertain. However, the latter is less likely: our 'zero effect mode' centres well around zero, without a large mass just above or below zero. Achieving this distribution would require participants to get this de-biasing 'just right', without overestimating the AMP effect and producing a negative AMP effect or underestimating it and leaving a positive AMP effect.*

(pseudo) randomization of their responses (i.e., without respect to the actual images), and that this does not explain the observed effects.[8]

Conclusions

All told, the resistance to the AMP effect observed here raises two major concerns for the use of AMP effect estimates as outcomes, both of them partially addressable. First, our results suggest that when participants have a strong incentive to conceal their true attitudes towards the prime, a subset of individuals appears able to manipulate their responses, which changes the distribution of AMP effects and significantly decreases the average AMP effect. While this downward bias cannot be entirely removed, investigators can better understand its potential magnitude by re-examining their data separately for individuals who report expending higher versus lower levels of effort to avoid the AMP effect. Second, as individuals who report expending higher levels of effort can more effectively misreport their attitudes, the AMP is susceptible to a problematic source of measurement error. If self-reported effort – whatever it measures – is correlated with other explanatory variables in a study using AMP data as an outcome, estimates of how these explanatory variables influence the outcome will be biased. We thus suggest that investigators (1) utilize a version of the AMP with the *Warning* language, as is now commonly employed, so they can then (2) collect data on self-reported effort following the AMP. Investigators can then test whether any explanatory variables thought to influence outcomes are correlated with self-reported effort.

In conclusion, we find that for a portion of the population, the AMP effect can be controlled – and essentially eliminated – when incentivized to avoid being influenced by the prime images. No special cover story or instruction on how to avoid being biased by the prime image is required. While the AMP remains a useful measure in cases where explicit reports are even more likely to be manipulated, investigators using the AMP should be aware of this risk of partial controllability, and particularly that it differs among individuals. To examine how problematic such variation is in a given experiment, investigators can examine the distributions of responses, rather than just the means, to see whether a subset of the sample showed an effect very near zero. They may also wish to obtain a self-reported measure of effort participants expend in avoiding the influence of the prime. This allows examination of AMP effects by level of effort, which can help the investigator to understand how differential controllability is influencing their results.

Acknowledgements

We thank Saleem Hussain for developing and managing the online java-based AMP experiment and Robert Pressel for excellent research assistance. For comments on earlier versions, we are grateful to Jay Fournier, Keith Payne, Spencer Piston, and Patricia Tan, as well as the editors of BJSP and two anonymous reviewers.

[8] *If participants seek to ignore the primes and targets and instead produce a sequence of responses intended to appear random, they often generate poorly randomized sequences (Chapanis, 1953). By characterizing the non-randomness of responses, we find evidence that such pseudo-randomization does sometimes occur, and when extreme enough, predicts an AMP effect near zero. However, the effect is not large enough to account for the effects of incentivization: those estimated to be 95% likely to have non-random responses have only an 11 percentage point smaller AMP effect than others. Moreover, those in the* incentivize *conditions were no more likely to show evidence of non-random response sequences than in other conditions. Of course, it remains possible that participants are skilled enough in creating random sequences to have evaded detection.*

References

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

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, *20*, 351–368. https://doi.org/10.1093/pan/mpr057

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. https://doi.org/10.1177/1745691610393980

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116. https://doi.org/10.1037//0022-3514.42.1.116

Chapanis, A. (1953). Random-number guessing behavior. *American Psychologist*, *8*, 1347–1363.

De Houwer, J., & Smith, C. T. (2013). Go with your gut! Effects in the Affect Misattribution Procedure become stronger when participants are encouraged to rely on their gut feelings. *Social Psychology*, *44*, 299–302. https://doi.org/10.1027/1864-9335/a000115

Eder, A. B., & Deutsch, R. (2015). Watch the target! Effects in the affective misattribution procedure become weaker (but not eliminated) when participants are motivated to provide accurate responses to the target. *Frontiers in Psychology*, *6*, 1–10. https://doi.org/10.3389/fpsyg.2015.01442

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. *Handbook of Research Methods in Social and Personality Psychology*, *2*, 283–310. https://doi.org/10.1017/CBO9780511996481.016

Gawronski, B., & Payne, B. K. (Eds.) (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: 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

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy*, *9*, 241–253. https://doi.org/10.1111/j.1530-2415.2009.01195.x

Hofmann, W., van Koningsbruggen, G. M., Stroebe, W., Ramanathan, S., & Aarts, H. (2010). As pleasure unfolds: Hedonic responses to tempting food. *Psychological Science*, *21*, 1863–1870. https://doi.org/10.1177/0956797610389186

Jarvis, W. B. G., & Petty, R. E. (1996). The Need to Evaluate. *Journal of Personality and Social Psychology, 70*, 172–194. https://doi.org/10.1037//0022-3514.70.1.172

Kalmoe, N., & Piston, S. (2013). Is implicit prejudice against blacks politically consequential? Evidence from the AMP. *Public Opinion Quarterly*, *77*, 305–322. https://doi.org/10.1093/poq/nfs051

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188. https://doi.org/10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon mechanical Turk. *Judgment and Decision Making*, *5*, 411–419. https://ssrn.com/abstract=1626226

Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, *39*, 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*, 277. https://doi.org/10.1037/0022-3514.89.3.277

Payne, B. K., Govorun, O., & Arbuckle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition & Emotion*, *22*, 238–271. https://doi.org/10.1080/02699930701357394

Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A process model of affect misattribution. *Personality and Social Psychology Bulletin*, *6*, 1397–1408. https://doi.org/10.1177/0146167210383440

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*, 367–374. https://doi.org/10.1016/j.jesp.2009.11.001

Payne, B. K., McClernon, F. J., & Dobbins, I. G. (2007). Automatic affective responses to smoking cues. *Experimental and Clinical Psychopharmacology*, *15*, 400–409. https://doi.org/10.1016/j.jesp.2009.11.001

Schreiber, F., Bohn, C., Aderka, I. M., Stangier, U., & Steil, R. (2012). Discrepancies between implicit and explicit self-esteem among adolescents with social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*, 1074–1081. https://doi.org/10.1016/j.jbtep.2012.05.003

Teige-Mocigemba, S., Penzl, B., Becker, M., Henn, L., & Klauer, K. C. (2015). Controlling the "uncontrollable": Faking effects on the affect misattribution procedure. *Cognition and Emotion*, *30*, 1470–1484. https://doi.org/10.1080/02699931.2015.1070793

Received 7 January 2017; revised version received 13 August 2017