Revisiting Dissociation Hypotheses with a Structural fit Approach: The Case of the Prepared Reflex Framework

Article *in* Journal of Experimental Social Psychology · January 2022

DOI: 10.1016/j.jesp.2022.104297

4 authors:

Jérémy Béna
Université Catholique de Louvain - UCLouvain
8 PUBLICATIONS **17** CITATIONS
SEE PROFILE

David Melnikoff
Northeastern University
13 PUBLICATIONS **352** CITATIONS
SEE PROFILE

Adrien Mierop
Université Catholique de Louvain - UCLouvain
24 PUBLICATIONS **224** CITATIONS
SEE PROFILE

Olivier Corneille
Université Catholique de Louvain - UCLouvain
152 PUBLICATIONS **4,916** CITATIONS
SEE PROFILE

Revisiting Dissociation Hypotheses with a Structural fit Approach: The Case of the Prepared Reflex Framework

Jérémy Béna[a], David Melnikoff[b], Adrien Mierop[a], & Olivier Corneille[a]

Manuscript accepted at *Journal of Experimental Social Psychology*. The final version may differ from this version.

doi: 10.1016/j.jesp.2022.104297

[a]UCLouvain, Belgium

10 Place du Cardinal Mercier, 1348, Louvain-la-Neuve, Belgium

[b]Department of Psychology, Northeastern University, Boston, MA, USA

360 Huntington Avenue, 118 Nightingale Hall, MA 02115, USA

Author Note

This work was supported by an FRS-FNRS grant [grant T.0061.18] awarded to Olivier Corneille.

Conflict of interest statement: The authors declare no conflict of interest.

Correspondence concerning this article should be addressed to Jérémy Béna, UCLouvain, PSP IPSY, 10 Place du Cardinal Mercier, 1348, Louvain-la-Neuve, Belgium. Email: jeremy.bena@uclouvain.be

Highlights

- Dissociation hypotheses are critical to implicit social cognition research.
- They are tested by comparing performance on “implicit” vs. “explicit” tasks.
- When tasks are structurally unfitted, no clear interpretation can be made.
- For the first time, we revisited a dissociation hypothesis using a structural fit approach.
- The dissociation hypothesis, otherwise supported, was rejected.

Abstract

Attitude and social cognition research often tests dissociations in performance on "explicit" and "implicit" measures using tasks that widely differ from each other. This prevents a clear interpretation of the findings. A *structural fit* approach, involving tasks that differ only on the factor of theoretical interest, should be preferred. Here, we revisited a dissociation hypothesis using a structural fit approach for the first time. Specifically, Melnikoff, Lambert, and Bargh's (2020) Prepared Reflex framework posits that the establishment and activation of an association between a planned action and its target make the valence of the action unintentionally spread to the target. As a result, the target is evaluated according to the valence of the action. Once the action plan is inactive, however, unintentional measures of attitudes should no longer reflect the valence of the planned action, as the association is not active anymore. In contrast, the valence of the inactive plan should continue to influence intentional measures of attitudes. Melnikoff et al. (2020) found support for this dissociation hypothesis in studies comparing evaluations on self-reports versus an Affect Misattribution Procedure (AMP): Whereas both measures captured the valence of the planned action when the plan was active, only the self-reports captured this valence after the plan was inactive. Here, contrary to the dissociation hypothesis, the plan valence influenced target evaluations on an intentional AMP when the plan was active but not when it was made inactive. These findings highlight the importance of using structurally-fitted tasks in attitude and social cognition research.

Keywords: Action control; Affect Misattribution Procedure; Attitudes; Automaticity; Dissociation; Prepared reflex.

Revisiting Dissociation Hypotheses with a Structural fit Approach: The Case of the Prepared Reflex Framework

Dissociations in performance on implicit and explicit tasks are critical to attitude and social cognition research. The implicit-explicit distinction can be interpreted in three ways (see Corneille & Hütter, 2020). A first perspective (i.e., "implicit-as-indirect") refers to the difference in performance on direct and indirect measures. A second perspective (i.e., "implicit-as-automatic") opposes performance on measures that vary in automaticity (e.g., awareness, control, intentionality, efficiency) (see Bargh, 1994; Melnikoff & Bargh, 2018). Finally, a third perspective (i.e., "implicit-as-associative") posits that more automatic measures capture associative processes (e.g., Kurdi & Banaji, 2017; McConnell & Rydell, 2014). The last two perspectives on the "implicit" construct are particularly relevant to theorization in psychological science. For instance, mental theories may predict that less information is retrieved from memory when participants respond under time pressure (i.e., for more automatic judgments), resulting in a divergence of outcomes on speeded vs. non-speeded tasks (e.g., the MODE model, Fazio, 1990; Ranganath et al., 2008). Likewise, associative theories may consider that some tasks (e.g., the Implicit Association Test; IAT) better capture attitudes that were acquired through associative processes (e.g., the Associative-Propositional Evaluation model, APE; Gawronski & Bodenhausen, 2006, 2011, 2014, 2018; the systems of evaluation model, McConnell & Rydell, 2014; Rydell & McConnell, 2006).

In the vast majority of studies, dissociations are examined by comparing performance on tasks that widely differ between each other (e.g., a self-report vs. an IAT). This makes it difficult to identify which features may be responsible for the observed dissociation (see also below). Two solutions have been proposed to overcome this problem. The first solution (Process Dissociation procedures) consists in estimating the contribution of mental processes

to task performance *within* a single task (for a recent review in social psychology, see e.g., Hütter & Klauer, 2016; see also Payne, 2008; Payne & Bishara, 2009). The second solution (Structural fit procedures) consists in comparing performance on tasks that vary only on the feature of interest (Payne et al., 2008). Surprisingly enough, we know of no published study that revisited a dissociation hypothesis by relying on a structural fit approach.

In the present research, we followed-up on Payne et al.'s (2008) recommendation to rely on a structural fit approach to revisit a dissociation between implicit/unintentional and explicit/intentional measures. Specifically, we revisited the dissociation hypothesis that Melnikoff, Lambert, and Bargh (2020) derived from their prepared reflex framework. In doing so, the present research proceeded to a more rigorous test of the prepared reflex framework's dissociation hypothesis and delivered insightful information on the value of the structural fit approach. The present research was implemented through an adversarial collaboration approach to offer the best guarantees regarding the diagnosticity of the test and the quality of the conclusions. Below, we outline Melnikoff et al.'s (2020) prepared reflex framework, and importantly their dissociation hypothesis that we revisited in two experiments using a structural fit approach.

To our knowledge, this is the first time that a dissociation hypothesis is revisited using a structural fit approach. This represents the main and higher-order goal of the current research. As a result, we will not elaborate on the details of the prepared reflex framework in the introduction by discussing, e.g., how competing models may account for the original effects. More generally, we note here that the structural fit recommendation made by Payne et al. (2008) has been rarely implemented.

Melnikoff et al.'s (2020) prepared reflex framework.

In a recent article, Melnikoff et al. (2020) proposed a new approach to attitude formation and change. Their prepared reflex framework draws on theories of action control

(e.g., Gollwitzer, 1999; Hommel, 2000; Hommel & Wiers, 2017), which have proved influential in psychological science but have been overlooked in attitude research. For a complete presentation of the framework, we invite the readers to refer to Melnikoff et al. We outline below the gist and main predictions of the framework.

Action control is thought to involve mental structures called *prepared reflexes*: stimulus-actions associations that form in working memory when people plan to carry out an action to a target. Melnikoff et al. (2020) proposed that prepared reflexes are a unique mechanism in evaluative processing, in addition to other mechanisms. The idea is that once a prepared reflex is formed, the valence of the planned action can be activated unintentionally on perception of the stimulus, thereby altering evaluative responses to the stimulus in the direction of the valence of the response. For instance, if Mary plans to treat Jane to dinner (a positive planned action), she may spontaneously evaluate Jane (the stimulus associated with the planned action) more positively due to the establishment and activation of a prepared reflex associating Jane with a positive action. When Mary perceives Jane, the associated positive planned action will unintentionally spring to mind along with its positive valence, resulting in a relatively positive response.

Melnikoff and colleagues (2020) derived several predictions from this framework that alternative models of attitude formation do not easily accommodate. Before introducing the dissociation hypothesis that is critical to the present research endeavor, we briefly outline the main hypotheses of the model. Following the *inaction* hypothesis, prepared reflexes change attitudes even if their corresponding action plans are never performed. For instance, even if Mary never treats Jane to dinner, her prepared reflex should cause her to evaluate Jane more positively — merely planning the positive action is sufficient. Another prediction is the *transience* hypothesis: prepared reflexes change attitudes only for as long as their underlying action plans are in place. Once action plans are inactive (e.g., terminated), effects on attitudes

are eliminated. Mary's prepared reflex would cause her to evaluate Jane more positively only for as long she plans to perform the positive action. If Mary were to abandon her plan, her original attitude toward Jane would be reinstated. The prepared reflex framework also predicts that effects of prepared reflexes on attitudes are unconstrained by the features of the stimuli being evaluated. The amount of attitude change elicited by a prepared reflex is a function of the valence of the planned action and the strength of its connection to the target stimulus — the content of the stimulus is irrelevant. The stimulus could be your best friend, a casual acquittance, or a mass murderer, and the effect of a prepared reflex on your evaluation of the target would be same. This is the *additivity hypothesis*.

Across six experiments, Melnikoff et al. (2020) obtained evidence for all three hypotheses, lending support to the prepared reflex framework while ruling out alternative mechanisms of attitude change, including inferential reasoning, cognitive dissonance, and biased scanning. Of note, support for the prepared reflex framework does not uniquely corroborate this model, as the effects Melnikoff et al. found are compatible with dual-process models of attitude learning (e.g., the APE model, Gawronski & Bodenhausen, 2006, 2018) and propositional models (e.g., De Houwer, 2018). The prepared reflex framework is compatible with such theorizations but independent from them as it is agnostic to the distinction between associative and propositional processes.

Melnikoff et al.'s (2020) dissociation hypothesis.

Importantly, the current research focuses on a fourth prediction of the prepared reflex framework: the *dissociation hypothesis*, which states that prepared reflexes should elicit different patterns of unintentionally expressed and intentionally expressed attitude change. The dissociation hypothesis begins with the observation that mental representations of stimuli and valence enter into different types of relations. One type of relation is *attributive*. A stimulus-valence relation is attributive if the valence is an attribute of the stimulus. For

instance, "Jane is good" is an attributive relation between the stimulus *Jane* and the valence *good*, because *good* is an attribute of *Jane*. Other relations between stimuli and valence are non-attributive, such as the relation between *Jane* and *good* in "I plan to do something good to Jane." Here, *good* is not an attribute of *Jane*, so their relation is non-attributive. As this example suggests, prepared reflexes create non-attributive relations between stimuli and valence.

Note that the concept of "non-attributive" should not be confused with the concept of "non-propositional," a term used in the attitudes literature to describe semantically unqualified connections between mental representations. Since the defining feature of attributive relations is their semantic content, non-propositional relations are non-attributive, but so are many propositional relations. The proposition "I plan to do something good to Jane" is one example. So, the claim that prepared reflexes create non-attributive relations implies nothing about the presence or absence of propositional content.

Melnikoff et al. (2020) 's dissociation hypothesis posits that non-attributive stimulus-valence relations directly alter unintentionally expressed attitudes but not intentionally expressed attitudes. The claim about unintentionally expressed attitudes is straightforward: any attributive or non-attributive relation between a stimulus and valence can mediate the spontaneous activation of valence on perception of the stimulus. Whether you think "Jane is good" or "I plan to do something good to Jane," you are apt to think *good* spontaneously when exposed to *Jane*. This process — the spontaneous activation of valence on stimulus perception — is all this is required to alter unintentionally expressed attitudes, so it follows that such attitudes can correspond to non-attributive stimulus-valence relations.

According to Melnikoff et al. (2020), this is not true of intentionally expressed attitudes, because these attitudes specifically describe attributive relations between stimuli and valence. "Jane is good" is an evaluation of Jane — "I plan to do something good to Jane"

is not. If intentionally expressed attitudes describe attributive relations, and prepared reflexes create non-attributive relations, it follows that prepared reflexes cannot change intentionally expressed attitudes on their own. A separate process is needed to convert the non-attributive relations created by prepared reflexes into the attributive relations that underlie such explicit attitudes. For instance, in order to alter intentionally expressed evaluations of Jane, the belief "I plan to do something good to Jane" must be converted into the belief "Jane is good." This could happen through processes such as inferential reasoning, misattribution, and the reduction of cognitive dissonance. What matters here is that effects of prepared reflexes on intentionally expressed attitudes are indirect, mediated by processes separate and distinct from the formation and operation of the prepared reflex itself. In contrast, effects of prepared reflexes on unintentionally expressed attitudes are direct, since they can correspond to non-attributive relations of the sort that prepared reflexes create.

Following this logic, Melnikoff et al. (2020) predicted that prepared reflexes would be associated with different patterns of unintentional and intentional attitude change. Consistent with this prediction, a dissociation emerged between performance on Affect Misattribution Procedure (AMP; Payne et al., 2005; Payne & Lundberg, 2014) measures and self-reports. The AMP measure adhered to the transience hypothesis but the self-report did not: when participants abandoned their action plans, their unintentional evaluations returned to their original, pre-plan state. In contrast, their intentional evaluation remained in their altered post-plan state.

A structural fit test of the dissociation hypothesis.

As mentioned above, however, comparisons between evaluative ratings and AMP outcomes do not allow for strong dissociation tests. This is because when a set of structural features vary between two tasks, one cannot attribute an observed effect to any specific feature. This important limitation is widespread in implicit social cognition research (for

recent discussions, see Corneille & Hütter, 2020; Gawronski, 2019; Hütter & Klauer, 2016; see also Payne, 2008; Payne et al., 2008), and also applies to the experiments by Melnikoff et al. (2020). The AMP and the evaluative self-reports did not just vary regarding how intentionally target evaluations were provided. They also varied regarding, for instance, (1) the relativity of the evaluations (they were relative in the AMP but absolute in self-reports), (2) the number of measures (there were 60 trials in the AMP, but three items in the self-reports), (3) the scale used (dichotomous responses in the AMP; 7-point Likert scales in the self-reports), (4) the use of pictures (AMP) versus verbal statements (self-reports). Any of these (or others) differences could be responsible for the dissociations found in Melnikoff et al. (2020) – not just differences in intentionality. Besides interfering with a precise interpretation of differences in task outcomes, such differences may additionally imply differences in measurement error and, as a result, in the general sensitivity of the tasks.

To overcome these interpretational ambiguities, we proceeded to a more stringent test of the dissociation hypothesis for transience. This was achieved in an "adversarial" collaboration spirit, by relying on a *modified* version of the AMP. This modified AMP was designed by Payne et al. (2008; for implementations of this task, see also Van Dessel et al., 2020; Zerhouni et al., 2018, 2019) to keep task structure as close as possible to the original, standard AMP, but by isolating the feature of interest: the intentionality of the evaluation. In the standard AMP, participants are primed by the target stimuli before evaluative classifications of Chinese characters. Because participants are explicitly instructed *not* to be influenced by the target stimuli and to draw on their feelings exclusively from the Chinese characters in reporting their evaluations, the obtained evaluations of the target stimuli can be seen as unintentional. In contrast, in the modified AMP, participants are instructed to draw on their feelings about the target stimuli and not to be influenced by the Chinese character. In this case, their evaluations of the target stimuli are more intentional (see e.g., Payne et al.,

2008; Van Dessel et al., 2020). To our knowledge, the present study is the first one that relied on a structural fit approach to test or revisit a dissociation hypothesis.

Overview of the experiments

Experiment 1 examines effects of Plan Valence (positive or negative) and Plan Status ("active" or "inactive") on the modified (i.e., more intentional) version of the AMP. Of main interest here is whether effects on the AMP will be sensitive to a Plan Valence manipulation at each Plan Status level. Should Plan valence influence target evaluations when the plan is active but not when the plan is inactive, this would contradict the dissociation hypothesis. This is because the critical effect would be found on a measure supposed to reflect attributive relations rather than non-attributive relations. Experiment 2 addresses significant departures from the original paradigm that were noticed by the first author of the original manuscript upon discussion of Experiment 1. It also relies on more stringent criteria for the selection of participants and on a more comprehensive analytic strategy. Furthermore, Experiment 2 involves measures on both the modified and the standard AMP. This second experiment allows examining whether the interaction between Plan Valence and Plan Status expected on the standard AMP extends to its intentional, modified AMP, counterpart.

Experiment 1

Because the modified AMP involves intentional evaluations of the target of the planned action, the dissociation hypothesis predicts that evaluations of the target in this task should be sensitive to Plan Valence under *both* levels of the Plan Status factor. That is, more positive evaluations of the target should be observed when the plan is positive than negative, and this should be the case whether the plan is active or inactive.

Alternatively, finding a significant effect of Plan Valence when the plan is active but not when it is inactive would contradict the dissociation hypothesis. This is because effects predicted exclusively for unintentional evaluations would now be observed for intentional evaluations. Finding a Plan Valence effect only when the plan is inactive would also lead to reject the dissociation hypothesis. Finally, finding no effect of Plan Valence on either level of Plan Status would speak to the lack of generalization of previous findings.

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The pre-registration, program script, and raw data are available on the Open Science Framework (https://osf.io/fa4gz/).

Participants and design

The design was a 2 (Plan Valence: positive vs. negative) × 2 (Plan Status: active vs. inactive) × 2 (Time: before plan induction vs. after plan induction), with the last factor within participants.

We recruited 215 English speaking participants living in the United States (our targeted sample size) on Prolific (40.93% female; M_{age} = 36.8; SD_{age} = 10.96). Participants were paid US$1.25 for completing the study. During the study, participants were told that the 10 best attorneys would win a US$10 bonus. After they learned that they would not play the game, participants were told that 10 randomly selected participants would get the bonus. Using the preregistered exclusion criteria, we excluded 34 participants (15.81% of the total, initial sample)[1]. This resulted in a final sample size of 181 participants.

[1] We pre-registered that participants would be excluded if they reported recognizing the Chinese characters. However, we were not able to apply this criterion, as we did not ask participants if they recognized some of the Chinese characters. Experiment 2 resolves this issue. Seven participants did not complete the study; 16 participants failed to correctly indicate whether they were instructed to help or harm the target; 11 participants had a mean response time under 200ms on either of the two AMPs.

To determine sample size, we relied on the effect size found by Melnikoff et al. (2020, Study 1). They found that the effect of Plan Valence on the *standard* AMP scores in the inactive condition was Cohen's $d = 0.72$ (after baseline correction). We performed the power analysis for *Cohen's d* = 0.50 for a safe test, aiming for a 95% statistical power in a unilateral independent samples *t*-test[2] (because of the unilateral hypothesis that positive planned actions would increase liking for the target compared to negative planned actions). The analysis indicated that a total sample of 176 participants was required to reach $1 - \beta = .95$ in the two levels of the Plan Status factor (i.e., 88 participants were required at each level). We aimed at collecting data on 215 participants to accommodate data exclusion (about 20% in Melnikoff et al., 2020) and potential data loss.

Materials and procedure

We programmed the experiment with OpenSesame (Mathôt et al., 2012). We used OSWeb (Mathôt & March, 2021) and JATOS (Lange et al., 2015) to run the study online on Prolific. The study was adapted from Melnikoff et al.'s (2020) Study 1. Six changes were applied to the original study. First, we did not manipulate stimulus valence (operationalized as the guilt or innocence of the target person). This manipulation is irrelevant to the question at hand. This is because the stimulus valence manipulation is related to the test of the prepared reflex's additivity hypothesis (i.e., evaluations should go in the direction of the planned action valence independent of the stimulus valence, as Melnikoff et al. [2020, Study

[2] Note, however, that this a priori power analysis overestimates the actual statistical power of our pre-registered analyses. This is because we planned to conduct separate independent samples *t*-tests in *each* Plan Status condition – doing so, we would need $n = 88$ at each Plan Valence*Plan Status level (not at each Plan Status level) to achieve the targeted power. Sensitivity power analyses showed that with $n = 44$ at each Plan Valence*Plan Status level, we have a 95% statistical power to detect effects as small as $d = 0.71$, which is very close to the first, non-conservative estimate of our targeted effect size. We still have an 80% power to detect effects as small ad $d = 0.53$, which is close to our more conservative estimate of the targeted effect size. In additional, non-preregistered analyses, we also relied on the full dataset, allowing for higher-powered tests of the focal hypothesis.

3] found using, e.g., Adolf Hitler as an extreme attitudinal object). The target was positive (i.e., innocent) in all conditions. Second, we adapted the standard AMP instructions for the modified AMP, because we wanted participants to intentionally evaluate the persons and not the Chinese characters. Third, there were less trials with the target picture in our modified AMP (10 target trials out of 60 critical trials, compared to 30 target trials in Melnikoff et al.), and more trials with control pictures (50 out of 60, compared 30 in Melnikoff et al., see below). Fourth, Francis West backstory was shorter than the original. Fifth, we removed most of the attention checks Melnikoff et al. used throughout their study. Sixth, we used a white background in the modified AMP, while Melnikoff et al. used a black background. Of note, the last four changes may be considered significant departures from the original procedure. This limitation is resolved in Experiment 2.

Upon clicking on the study link, participants were told that we were interested in people's ability to ignore goal-relevant imagery. We explained that they would complete a series of visual perception tasks (the modified AMPs), and that they would play a game called "Attorney at Law", based on the trial of a man accused of murder, Francis West.

Right after this introduction, participants learned about Francis West, the accused man. Participants saw a picture of Francis West and read how he got accused: Francis West was at the beach with his best friend Roger. Roger got pulled under water by a strong current, and Francis risked his life to save him – without success, resulting in Roger's death.

Next, participants performed the modified AMP for the first time (i.e., baseline Time1 AMP, before plan induction). Participants were instructed that they would see pairs of pictures briefly displayed one after the other, where the first picture shows a person and the second shows a Chinese character. Their task, participants read, was to judge whether each person is a pleasant or an unpleasant individual. To mimic the standard AMP, participants were further told that the Chinese characters can sometimes bias people's judgement of the

individuals, and to try their absolute best not to let the Chinese characters influence their judgments. This first modified AMP consisted of 10 practice control trials (not analyzed) and 60 critical trials (10 target trials, 50 control trials). Target trials began with the picture of Francis West that participants saw when they were introduced to Francis. Control trials began with a picture of a middle-aged, white male – randomly selected from five pictures (all displayed 10 times each in total in critical trials). Each trial consisted of a person picture displayed for 75ms, a Chinese character displayed for 100ms, and a backward mask (a white noise image). The backward mask remained onscreen until participants provided a "Pleasant" (key "E") or "Unpleasant" (key "I") response.

Once the modified AMP completed, participants were randomly assigned to the role of defense attorney (positive Plan Valence) or prosecuting attorney (negative Plan Valence). Participants in the role of defense attorney were requested to help Francis, by presenting the jury with as much positive, exonerating, evidence as possible. Participants in the role of prosecuting attorney were requested to harm Francis by presenting the jury with as much negative, incriminating evidence as possible. To motivate participants, they were informed that US$10 would be rewarded to the 10 best attorneys. To ensure participants formed the correct planned action, they were then asked to say out loud "If I see Francis West, then I will eliminate the word "GUILTY" (positive Plan Valence)/"INNOCENT" (negative Plan Valence) as fast as possible!" three times before continuing. They were further asked to report if they were told to prosecute or to defend Francis (participants who incorrectly responded were excluded, see above).

We then implemented the Plan Status manipulation. Participants were randomly assigned to either an "active" condition or to an "inactive" plan condition. In the "inactive" condition, participants were told that they would not play Attorney at Law, after all, as the game is – they were told, not compatible with their operating system. Participants were

further told that they were still eligible for a US$10 bonus. Participants in the "active" condition only saw this message right before the debriefing. In the end, no participants actually played "Attorney at Law," as the game was part of the study's cover story.

Next, all participants performed the modified AMP for the second time (i.e., Time2 AMP, after plan induction). The procedure for this AMP differed from the first one only in that (1) the 10 practice trials were absent, and (2) the instructions were reduced in length (as participants were already familiarized with the task).

Finally, participants used 7-point Likert scales to report how much they (dis)like Francis (1: "Strongly dislike"; 7: "Strongly like"), how positively/negatively they feel toward Francis (1: "Very negatively"; 7: "Very positively"), and how good/bad Francis is (1: "Very bad"; 7: "Very good").

Results

We performed all analyses in R (R Core Team, 2020). We used the R packages *BayesFactor* (Morey & Rouder, 2018, version 0.9.12-4.2) to compute the default Bayes independent sample *t*-test (Rouder et al., 2009), *effsize* (Torchiano, 2020, version 0.8.1) to compute the Cohen's *d*, *ggplot*2 (Wickham, 2016) and *ggpubr* (Kassambara, 2020, version 0.4.0) to make the raincloud plots (Allen et al., 2021).

In each modified AMP, the scores are the proportions of "pleasant" responses in target trials (i.e., with a picture of Francis West) minus the proportions of "pleasant" responses in control trials (i.e., with control pictures). Zero-scores indicate identical evaluations for Francis West and controls, negative scores indicate more "pleasant" responses for controls over Francis West, and positive scores indicate more "pleasant" responses for Francis West over controls.

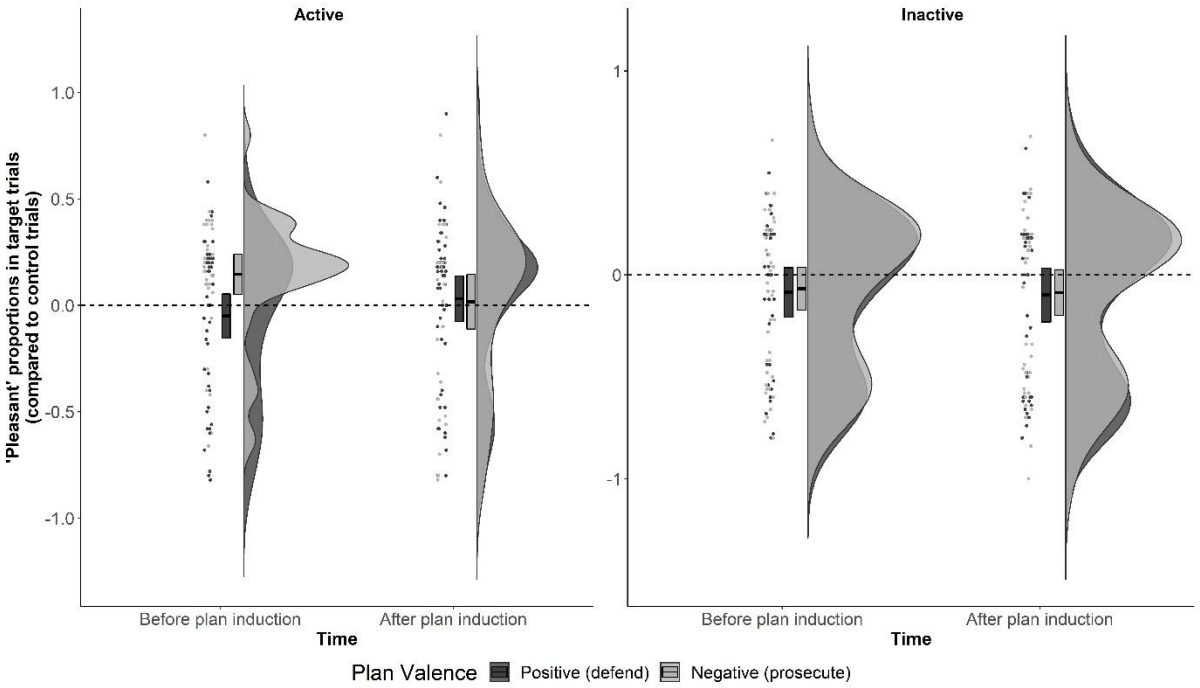

Figure 1. Proportions of "pleasant" responses in the modified AMP in target trials (minus proportions of "pleasant" responses in control trials) as a function of Time, Plan Valence, and Plan Status (left: active; right: inactive). The dots are the participants scores (jittered). The lower and upper limits of the boxplots are the 95% confidence intervals, with the mean in between. The distributions represent the kernel probability density of the datas in each Time × Plan valence condition. Dashed horizonal line: no preference between the proportions of "pleasant" responses between the target trials and the control trials.

Because the internal reliability of the three evaluative self-rating items was high (Cronbach's $\alpha = .93$), the self-reported evaluative rating scores are the participants' average response across the three items (the higher the scores, the more positive the evaluations).

For the modified AMP, we report the preregistered analyses and additional non-preregistered analyses separately. For the self-reported evaluative ratings, we report only additional exploratory analyses, as we did not preregister analyses.

Modified AMP

As preregistered, we tested whether the modified AMP scores differed as a function of Plan Valence, separately in each Plan Status condition. For each participant, we computed the final modified AMP scores as the difference between scores obtained in the Time2 AMP and

scores obtained in the Time1 AMP. This is conceptually similar to modified AMP scores after controlling for baseline AMP scores. The main results are displayed in Figure 1.

When the plan was active, modified AMP scores were more positive after participants formed a positive plan (M = .08; SD = .25) than after they formed a negative plan (M = -.13; SD = .32), $t(87) = 3.43$, $p = .0005$ (one-sided), $d = 0.735$. This difference was not significant anymore when the plan was inactive, $t(90) = 0.13$, $p = .447$ (one-sided), $d = 0.028$.

We performed a default Bayesian (two-sided) independent samples t-test (Cauchy prior = .707) to estimate support for the null hypothesis (no effect of Plan Valence in the inactive condition) against the alternative hypothesis of an effect (Bayes Factor noted as BF_{01}). We found moderate evidence for the null hypothesis, $BF_{01} = 4.5 \pm 0.03\%$. However, the Bayesian analysis yields only insufficient support for the hypothesis of no effect, as the Bayes Factor is below our preregistered cut-off value (i.e., $BF_{01} = 5$).

Additional (non-preregistered) analyses

In the preregistered analyses, we found a significant effect of Plan Valence in the active condition but not in the inactive condition. However, this difference might itself be non-significant. To provide a more complete, higher-powered picture of the data, we conducted a 2 (Plan Valence: positive vs. negative) × 2 (Plan Status: active vs. inactive) × 2 (Time: before plan induction vs. after plan induction) mixed ANOVA on the AMP scores to estimate the three-way interaction between Plan Valence, Plan Status, and Time. This interaction was significant, $F(1, 177) = 7.03$, $p = .009$, $\eta^2_G = .005$.

We tested the simple effects of the Plan Valence × Time interaction at each Plan Status level using non-preregistered, non-adjusted pairwise multiple comparisons (computed based on the full model). We replicated the pattern we found in the preregistered analyses. No effect was significant when the plan was inactive (larger effect: $t(177) = 0.54$, $p = .587$).

When the plan was active, participants who formed a positive plan had more positive AMP scores after plan induction than before, $t(177) = -2.26$, $p = .025$. The reverse was true for participants who formed a negative plan: they had more negative AMP scores after plan induction than before, $t(177) = 3.11$, $p = .002$. Note that at Time1, AMP scores were higher for participants who formed a negative plan compared to a positive plan, $t(222) = -2.42$, $p = .016$. This result suggests a failure of random assignment, as an effect of Plan Valence is evident before Plan Valence induction. As our analyses are either baseline-controlled (preregistered analyses) or involve Time as a factor (additional analyses), this effect does not impede our focal tests.

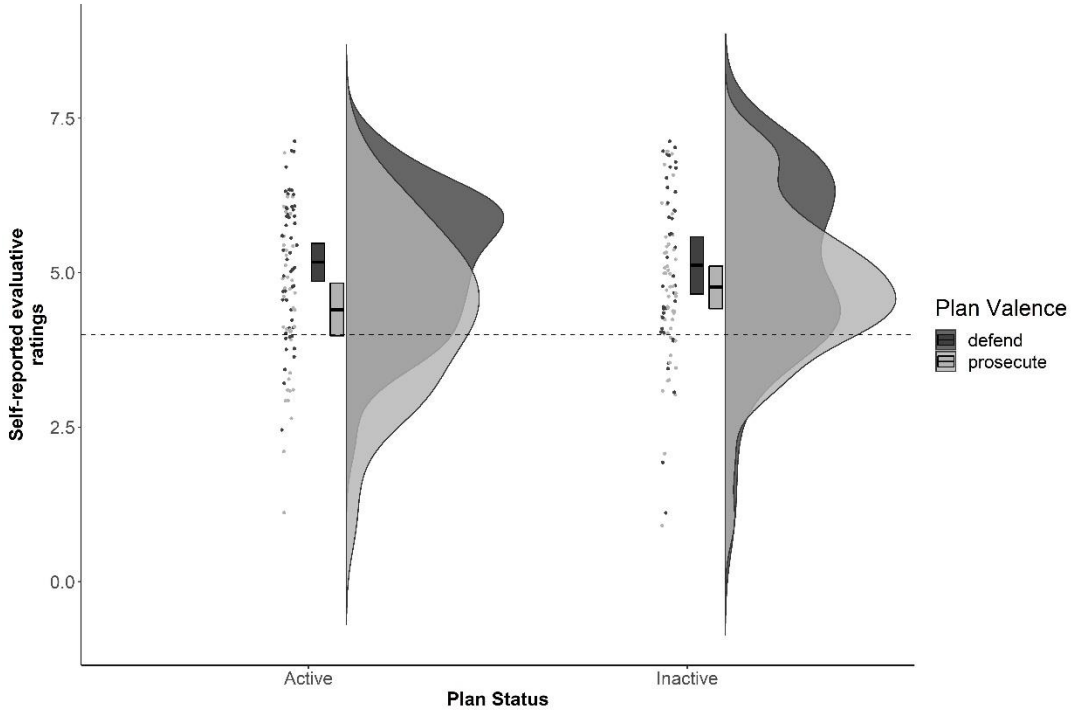

Figure 2. Self-reported evaluative ratings of the target (Francis West) as a function of Plan Valence and Plan Status. The dots are the participants scores (jittered). The lower and upper limits of the boxplots are the 95% confidence intervals, with the mean in between. The distributions represent the kernel probability density of the data in each Plan Valence × Plan Status condition. Dashed horizonal line: neutral ratings (Francis West neither evaluated positively or negatively).

Self-ratings: Non-preregistered analyses

We did not preregister analyses on the self-ratings, but because it is yet another direct (intentional) evaluative measure, we report results on this task for exploratory purposes (analyses of the self-ratings were preregistered in Experiment 2).

We conducted a 2 (Plan Valence) × 2 (Plan Status) between-subjects ANOVA on the self-reported evaluative ratings (see Figure 2). The main effect of Plan Valence was significant, $F(1, 177) = 8.76$, $p = .004$, $\eta^2_G = .047$: Participants who formed a positive plan gave more positive ratings ($M = 5.15$; $SD = 1.25$) than participants who formed a negative plan ($M = 4.61$; $SD = 1.27$). This Plan Valence effect was not qualified by Plan Status, $F(1, 177) = 1.19$, $p = .277$, $\eta^2_G = .007$. The main effect of Plan Status was not significant either, $F(1, 177) = 0.65$, $p = .42$, $\eta^2_G = .004$.

Because we are interested in the pattern of simple effects under each Plan Status condition, we decomposed the two-way interaction between Plan Valence and plan status, even if it is not statistically significant. We found a pattern similar to the one obtained on the modified AMP scores (see Figure 2). When the plan was active, self-ratings were more positive when participants formed a positive plan than when they formed a negative plan, $t(177) = 2.84$, $p = .005$. This difference was not significant when the plan was inactive, $t(177) = 1.33$, $p = .18$.

Discussion

In Experiment 1, we found an effect of plan valence on both the modified AMP and the self-reported evaluations of the target. Critically, we found a simple effect of Plan Valence on both the modified AMP and evaluative ratings when the plan was active (i.e., participants intended to play the game), but not when the plan was inactive (i.e., participants no longer intended to play the game). The Bayesian analyses, however, yielded insufficient evidence for the hypothesis of no effect when the plan was inactive. Hence, evidence is mixed. In addition, as we pointed out in the study overview, Experiment 1 introduced several changes compared to Melnikoff et al.'s (2020, Study 1) original procedure. Of particular concern, the modified AMP involved fewer data points than the AMP used in the original studies.

Experiment 2

The many departures from the original study implemented in Experiment 1, as well as the ambiguous evidence collected, are undesirable. Experiment 2 was designed to overcome these limitations. First, we relied on a procedure that was much closer to Melnikoff et al. (2020, Study 1). Second, we had participants complete either the modified or the standard

AMP. Third, we preregistered a more comprehensive analytic strategy. Fourth, more stringent criteria were set for the selection of participants.

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The pre-registration, program scripts, data files, and analyses scripts are available on the Open Science Framework (https://osf.io/37msb/)[3].

Participants

The design was a 2 (AMP condition: standard vs. modified) × 2 (Plan Valence: positive vs. negative) × 2 (Plan Status: active vs. inactive) × 2 (Time: before plan induction vs. after plan induction), with the last factor within participants.

We recruited 460 participants (our targeted sample size) on Prolific. Using the same rationale as Melnikoff et al. (2020), we aimed for a statistical power of 80%[4] to detect a three-way interaction as small as $\eta^2_p = .02$ in mixed 2 × 2 × 2 ANOVAs (Plan Valence and Plan Status as between-subjects factors, and Time as a within-subjects factor with a correlation of $r = .30$ between the two repeated-measures). As we planned to run separate mixed ANOVAs on scores from each AMP condition (standard; modified), we needed 192

[3] We conducted another experiment with a similar design of the one reported here as Experiment 2. However, it suffered from the exact same limitations as Experient 1. In addition, this study was likely to be underpowered, as 215 participants in total were recruited for a 2 (AMP condition: standard vs. modified) × 2 (Plan Valence: positive vs. negative) × 2 (Plan Status: active vs. inactive) × 2 (Time: before plan induction vs. after plan induction, within participants) design. Therefore, we decided not to report the study. Its pre-registration, program script, and raw data are available on the Open Science Framework (https://osf.io/fa4gz/). Please also note that the introduction to Experiment 2 reported in the pre-registration was based on preliminary analyses of Experiment 1 that did not include baseline AMP measures, thereby introducing yet another important departure from Melnikoff et al. (2020).

[4] We aimed for less power (80%) in Experiment 2 than in Experiment 1 (95%) as the study would have been too costly otherwise (i.e., its duration was longer than Experiment 1 and it targeted a small-to-medium interactive effect rather than medium-to-large simple effects as was the case in Experiment 1).

participants in each AMP condition to achieve the targeted power. Anticipating an exclusion rate of approximately 20% (Melnikoff et al., 2020), we aimed for a total of 460 participants.

Participants (1) were English speakers, (2) declared to live in the United States, (3) never volunteered in a Prolific study using the "Attorney at Law" basic procedure, (4) had an approval rate of at least 95%, and (5) had at least 100 previous submissions. Participants were paid US$2.75 for completing the study. During the study, they were told that the 10 best attorneys would win a US$10 bonus. After they learned that they would not play the game, participants were told that 10 randomly selected participants would get the bonus.

Using the pre-registered exclusion criteria, we excluded 74 participants (16.09% of the total, initial sample)[5]. This resulted in a final sample size of 386 participants (54.92% female; M_{age} = 37.29; SD_{age} = 13.49). There were 186 participants in the standard AMP condition (56.99% female; M_{age} = 36.2; SD_{age} = 13.3; 83 participants in the active condition; 94 participants in the positive plan valence condition), and 200 participants in the modified AMP condition (53% female; M_{age} = 38.3; SD_{age} = 13.62; 109 participants in the active condition; 109 participants in the positive plan valence condition).

Materials and procedure

We programmed the experiment with Qualtrics and Minno.js (Bengayev, 2020; Zlotnick et al., 2015). The study closely resembles Melnikoff et al.'s (2020) Study 1, with four departures intentionally implemented for the purpose of this study.

[5] Twelve participants failed to correctly indicate whether they were instructed to help or harm the target; 14 participants reported to know the meaning of at least some of the Chinese characters; 34 participants failed to correctly indicate whether they were instructed to evaluate the Chinese characters (standard, direct AMP) or the Persons (modified, indirect AMP) in at least one of the two administered AMPs; 19 participants had a mean response time under 200ms on either of the two AMPs. Five participants were excluded based on more than one exclusion criterion.

First, similar to Experiment 1, we did not manipulate stimulus valence (operationalized as the guilt or innocence of the target person, Francis West). This is because this manipulation is irrelevant to the question at hand. The target stimulus (Francis West) was positive (innocent) in all conditions.

Second, we randomly allocated participants to perform either the standard AMP or the modified AMP (both before and after plan induction). Participants in the standard AMP condition were instructed to judge the Chinese characters while ignoring the Persons. Conversely, participants in the modified AMP condition were instructed to judge the Persons while ignoring the Chinese characters.

Third, at the end of the study, we asked participants to report whether they were instructed to judge the Persons or the Chinese characters. We added this question to ensure that participants conformed to the instructions they received in the (standard or modified) AMP. We used participants' responses to this question as an exclusion criterion (see above).

Fourth, participants were told that we were interested in the way people perceived goal-relevant visual imagery (not that we are interested in people's ability to ignore goal-relevant imagery). This is because participants in the modified AMP condition would not be asked to ignore the individuals in the AMP.

Results

We used R (R Core Team, 2020) for our main analyses. We conducted the frequentist analyses of variance (ANOVAs) with *afex* (Singmann et al., 2020, version 0.28-0), and the default Bayesian ANOVAs (Rouder et al., 2012) with *BayesFactor* (Morey & Rouder, 2018, version 0.9.12-4.2). We conducted pairwise multiple comparisons with *emmeans* (Lenth, 2020, version 1.5.2-1). We made the raincloud plots (Allen et al., 2021) with *ggplot*2 (Wickham, 2016) and *ggpubr* (Kassambara, 2020, version 0.4.0). Complementarily, we

conducted non-preregistered Bayesian analyses with empirical, informed priors. Because their results are similar to the default Bayesian ANOVAs reported in the main text, those analyses are not reported.

In the default Bayesian ANOVAs, we used the default medium r scale (1/2) for both fixed and random effects. As we were interested in the Bayes Factor (BF) for the interaction (a three-way interaction on the standard and modified AMP scores; a two-way interaction on self-reports), we computed the Bayes Factor of each model against the null hypothesis (of no effect). To obtain the BF specifically for the interactions (and not for the full model containing the interaction), we contrasted the performance of the full model including the relevant interaction to the performance of the model omitting the interaction. For example, a BF of 4 indicates that data are four times more likely under the model with the relevant interaction than under the model without it. A BF of .25 indicates that data are four times more likely under the model without the relevant interaction than under the model with it.

In each AMP, the scores are the proportions of "pleasant" responses in target trials (i.e., with a picture of Francis West) minus the proportions of "pleasant" responses in control trials (i.e., with control pictures). Negative scores indicate more "pleasant" responses for controls over Francis West, and positive scores indicate more "pleasant" responses for Francis West over controls.

Because the internal reliability of the three self-rating items was high (Cronbach's α = .93), the self-reported evaluative rating scores are the participants' average response across the three items (the higher the scores, the more positive the evaluations).

For each task (standard AMP, modified AMP, self-reported evaluative ratings), we report separately the preregistered analyses and additional, non-preregistered analyses.

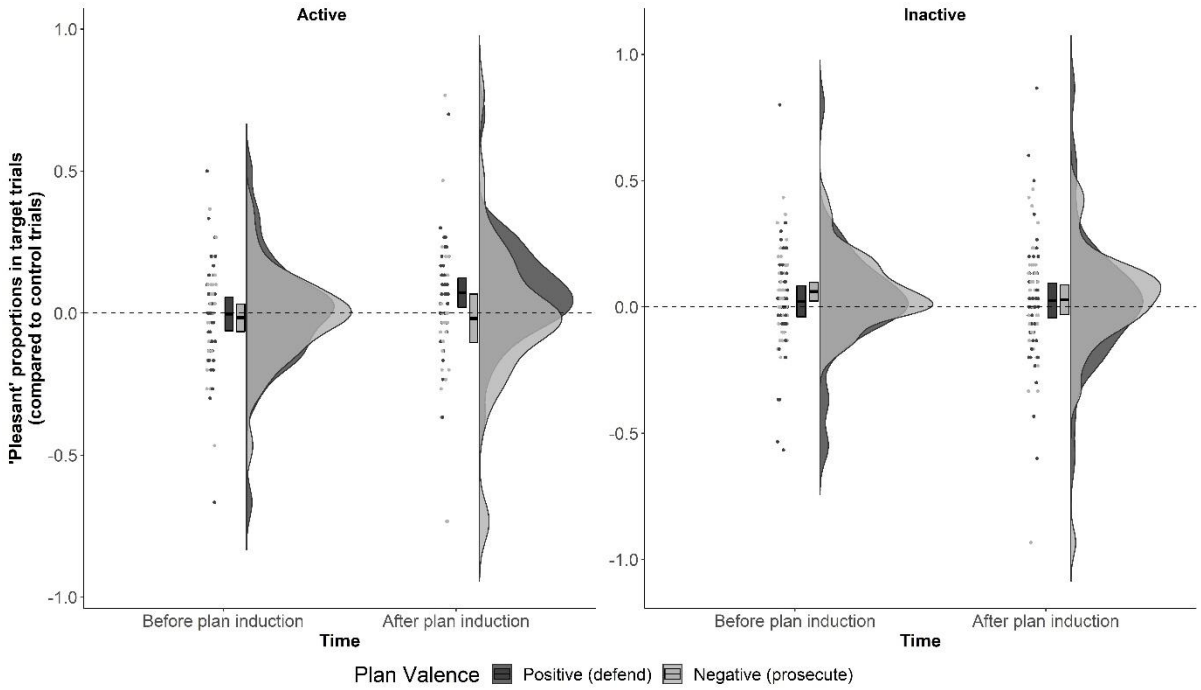

Figure 3. Proportions of "pleasant" responses in the *standard* AMP in target trials (minus proportions of "pleasant" responses in control trials) as a function of Time, Plan Valence, and Plan Status (left: active; right: inactive). The dots are the participants scores (jittered). The lower and upper limits of the boxplots are the 95% confidence intervals, with the mean in between. The distributions represent the kernel probability density of the data in each Time × Plan valence condition. Dashed horizonal line: no preference between the proportions of "pleasant" responses between the target trials and the control trials.

Standard AMP

Preregistered analyses

We conducted a 2 (Plan Valence: positive vs. negative) × 2 (Plan Status: active vs. inactive) × 2 (Time: before plan induction vs. after plan induction) mixed ANOVA on the standard AMP scores. The results are displayed in Figure 3. We found no significant main or interactive effect, $Fs(1, 182) \leq 2.98$, $ps \geq .086$, $\eta^2_G \leq .008$. Importantly, the three-way interaction was not statistically significant, $F(1, 182) = 0.43$, $p = .512$, $\eta^2_G \leq .001$. A Bayesian mixed ANOVA yielded moderate evidence *against* the three-way interaction, BF = 0.22 ± 24.47%.

Additional (non-preregistered) analyses

The sample size for the standard AMP condition was very close to the one that we had planned (i.e., 186 participants were retained in the final analysis instead of 192). However, it was smaller than that used in Melnikoff et al.'s (2020) studies (i.e., up to 371 participants, because Melnikoff et al. manipulated an additional factor, which is the valence of the target). Therefore, and also because we are interested in the simple effects (i.e., the effect of Plan Valence at each Plan Status level before and after plan induction), we decided to test the simple effects of the Plan Valence × Time interaction at each Plan Status level using non-preregistered, non-adjusted pairwise multiple comparisons (computed based on the full model).

No effect was significant when the plan was inactive (larger effect: $t(182) = 1.04$, $p = .3$). When the plan was active and after plan induction, AMP scores were more positive when participants formed a positive plan than when they formed a negative plan, $t(314) = 2.05$, $p = .041$. This effect was not significant before plan induction, $t(314) = .3$, $p = .765$. Participants who formed a positive plan had more positive AMP scores after plan induction than before, $t(182) = -2.26$, $p = .025$, but the effect of Time was not significant for participants who formed a negative plan, $t(182) = 0.07$, $p = .942$.

Modified AMP

Preregistered analyses

We conducted another 2 (Plan Valence) × 2 (Plan Status) × 2 (Time) mixed ANOVA, this time on the modified AMP scores (when participants were instructed to rate the primes, not the Chinese characters). The results are displayed in Figure 4. We found a significant main effect of Plan Valence, $F(1, 196) = 4.87$, $p = .028$, $\eta^2_G = .02$: AMP scores were more positive when participants formed a positive plan ($M = .17$; $SD = .26$) than when they formed a negative plan ($M = .08$; $SD = .35$). This Plan Valence effect was qualified by Time, $F(1,$

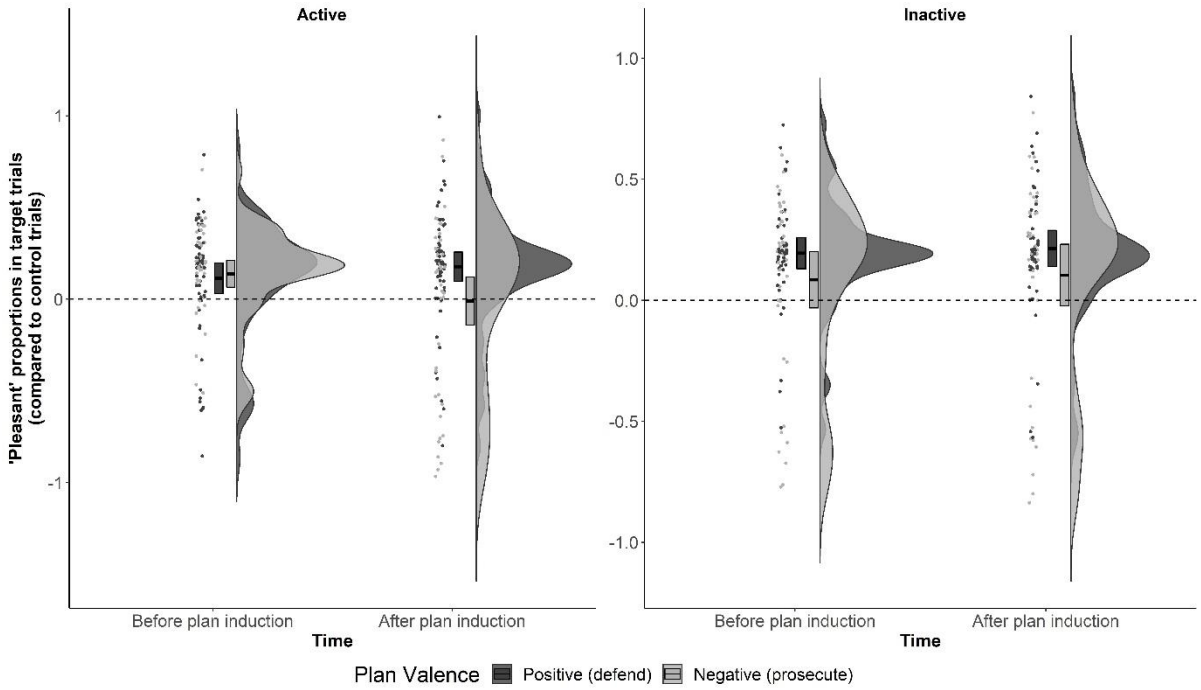

Figure 4. Proportions of "pleasant" responses in the *modified* AMP in target trials (minus proportions of "pleasant" responses in control trials) as a function of Time, Plan Valence, and Plan Status (left: active; right: inactive). The dots are the participants scores (jittered). The lower and upper limits of the boxplots are the 95% confidence intervals, with the mean in between. The distributions represent the kernel probability density of the data in each Time × Plan Valence condition. Dashed horizonal line: no preference between the proportions of "pleasant" responses between the target trials and the control trials.

196) = 7.8, $p = .006$, $\eta^2_G = .006$. Importantly, the three-way interaction was significant, $F(1, 196) = 7.7$, $p = .006$, $\eta^2_G = .006$. A Bayesian mixed ANOVA further yielded moderate evidence for the three-way interaction, BF = 5.6 ± 10.74%.

To decompose the three-way interaction, we conducted additional Plan Valence × Time mixed ANOVAs on the AMP scores in each Plan Status condition. Note, however, that these analyses should be interpreted with caution, as they only rely on portions of the data (we complement these analyses in the *Additional (non-preregistered) analyses* section below). When the plan was inactive, no effect was significant, $Fs(1, 89) \leq 2.94$, $ps \geq .09$, $\eta^2_G \leq .029$. In contrast, the Plan Valence × Time interaction was significant when the plan was active, $F(1, 107) = 12.11$, $p < .001$, $\eta^2_G = .024$. As displayed in Figure 4, the effect of Plan Valence was evident only after plan induction, when the plan was active: AMP scores were

more positive when participants formed a positive plan (M = .18; SD = .31) than when they formed a negative plan (M = -.01; SD = .46).

Additional (non-preregistered) analyses

As we just alluded to, a possible drawback of performing separate ANOVAs is that we lose information and power, as the analyses rely on subsets of the data. To overcome this concern and as was done on the standard AMP scores, we ran complementary non-preregistered pairwise multiple comparisons (computed based on the full model) to test the simple effects in each Plan Status condition.

Results are similar to the separate ANOVAs reported above. No effect was significant when the plan was inactive (larger effect: $t(269) = -1.85$, $p = .252$). When the plan was active, modified AMP scores were more positive when participants formed a positive plan than when they formed a negative plan after plan induction, $t(269) = 2.93$, $p = .019$. This effect was not significant before plan induction, $t(269) = -0.38$, $p = .98$. Participants who formed a negative plan had more positive AMP scores before plan induction than after, $t(196) = 3.91$, $p = .0007$, but the effect of Time was not significant for participants who formed a positive plan, $t(196) = -1.82$, $p = .266$.

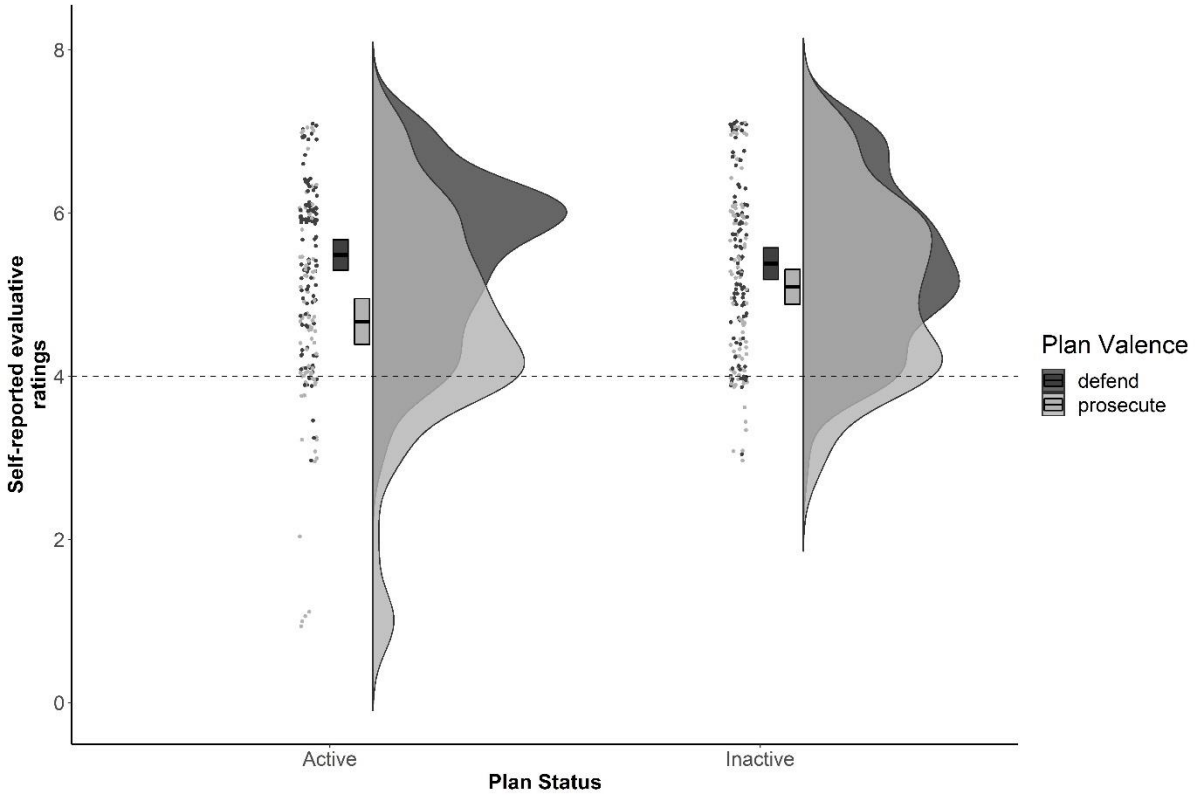

Figure 5. Self-reported evaluative ratings of the target (Francis West) as a function of Plan Valence and Plan Status. The dots are the participants scores (jittered). The lower and upper limits of the boxplots are the 95% confidence intervals, with the mean in between. The distributions represent the kernel probability density of the data in each Plan Valence × Plan Status condition. Dashed horizonal line: neutral ratings (Francis West neither evaluated positively or negatively).

Self-reported evaluative ratings

Preregistered analyses

We conducted a 2 (Plan Valence) × 2 (Plan Status) between-subjects ANOVA the self-reported evaluative ratings. The results are displayed in Figure 5. The main effect of Plan Valence was significant, $F(1, 382) = 25.07$, $p < .001$, $\eta^2_G = .062$. Participants who formed a positive plan gave more positive ratings ($M = 5.44$; $SD = 0.97$) than participants who formed a negative plan ($M = 4.89$; $SD = 1.2$). This Plan Valence effect was qualified by Plan Status, $F(1, 382) = 5.82$, $p = .016$, $\eta^2_G = .015$. The main effect of Plan Status was not significant, $F(1, 382) = 2.08$, $p = .15$, $\eta^2_G = .005$. A Bayesian between-subjects ANOVA yielded inconclusive evidence for or against the two-way interaction, BF = 2.52 ± 11.47%.

Additional (non-preregistered) analyses

Non-preregistered pairwise multiple comparisons in each Plan Status condition showed the expected Plan Valence effect in the active condition, $t(382) = 5.22$, $p < .0001$. The effect of Plan Valence effect was not statistically significant in the inactive condition, despite being close to the conventional alpha level, $t(382) = 1.84$, $p = .066$. Although it is difficult to conclude that the Plan Valence effect was cancelled in the inactive condition, the results suggest that the effect is at least stronger in the active condition.

Discussion

Consistent with Experiment 1 but inconsistent with the dissociation hypothesis, we found a simple effect of Plan Valence on the modified AMP when the plan was active, but not when the plan was inactive. The Bayesian analyses yielded evidence for the three-way interaction between time, plan valence, and plan status on the modified AMP, but evidence *against* this interaction on the standard AMP. In the standard AMP, we did not find the expected effect of plan valence, nor did we observe an interaction effect between plan valence and plan status, but we did observe the predicted pattern of simple main effects. Although less conclusive, the findings on the self-reported measure converged with those of the modified AMP. Here too, the plan valence influenced the target evaluation when the plan was active, but less clearly so when it was inactive.

General discussion

The present findings indicate the danger of relying on unfitted task comparison procedures when testing dissociation hypotheses. When effects differ on two structurally unrelated tasks (e.g., an AMP and a self-report; an Implicit Association Test and a feeling-thermometer), it is virtually impossible to identify which of the many structural differences between the tasks – or combination of these – are responsible for the observed difference in

effects. This is a major limitation when it comes to comparing and interpreting outcomes from so-called "implicit" and "explicit" measures (for recent discussions, see Corneille & Hütter, 2020; Gawronski, 2019; in memory research, see Roediger, 1990; Roediger, Weldon, Stadler, & Riegler, 1992).

In two preregistered experiments, we proceeded to a more rigorous test of the prepared reflex's (transience) dissociation hypothesis by relying on a structural fit approach. This allowed comparing performance on tasks that varied only on the factor of interest: intentionality. We additionally examined effects on self-reports and, for Experiment 2, on the standard AMP. Whereas Experiment 1 departed from the original procedures implemented by Melnikoff et al. (2020) in significant ways, Experiment 2 consisted of a close replication and extension of the original study. As we explain below, the use of the modified version of the AMP leads to rejecting the (transience) dissociation hypothesis of the prepared reflex framework, either in its mental model assumption or in its measurement assumption. In this general discussion, we discuss the implications of the current findings for the prepared reflex framework and for the study of dissociations.

The present research supports several predictions of the prepared reflex framework. In particular, it supports the inaction hypothesis (we found evaluative effects with a *planned* action alone) and the transience hypothesis (plan status did moderate the plan valence, although on the modified AMP and on the evaluative self-reports but not in the standard AMP). However, the findings lead to rejecting its dissociation hypothesis for transience. More specifically, they lead to rejecting this dissociation hypothesis *if* one assumes a sharp distinction between explicit/intentional tasks thought to capture attributive relations (e.g., a self-report, a modified AMP) vs. implicit/unintentional tasks thought to capture non-attributive relations (e.g., the standard AMP). The latter insight points to three levels of accuracy in the study of dissociations. By far, the weakest level of analysis is one that draws

dissociative conclusions from structurally unfitted task comparisons procedures. Unfortunately, this procedure is massively favored in social cognition and attitude research despite repeated notes of caution (see Gawronski, 2019; Gawronski et al., 2020; Payne et al., 2008). Then comes structural fit procedures like the one we implemented here. These procedures should be favored as they offer much stronger guarantees that comparisons are made across tasks that vary on the theoretical factor of interests (here: intentionality) rather than confounds. It is unfortunate that this approach recommended over one decade ago (Payne et al., 2008) has been overlooked. The highest level in our view is the one proposed by Process Dissociation procedures, which compare performance within a same task for conditions where the mental process(es) of interest operate antagonistically or synergistically (Hütter & Klauer, 2016; Jacoby, 1991; Yonelinas & Jacoby, 2012). Process Dissociation procedures come with their own assumptions, which can be occasionally violated (e.g., Klauer et al., 2015). However, provided an experimental validation of the parameters, to date they allow for the strongest possible tests of dissociative effects. As a side comment, the distinction between implicit and explicit *measures* makes little sense under this third approach (for a discussion, see Corneille & Hütter, 2020).

As we alluded to above, we found support for the transience hypothesis only on intentional measures but not on less intentional measures (the standard AMP). This last result is surprising because Melnikoff et al. (2020) consistently found that deactivating the planned action reduced its effect on unintentional measures to non-significance. In Experiment 2, we only found this effect on more intentional measures (the modified AMP and self-reports). As a result, the pattern we found suggests a dissociation opposite to the one expected under the prepared reflex framework: intentional measures might be more sensitive to deactivating a planned action than less intentional measures when a structural fit approach is used. Why this specific pattern of results occurred and why we did not replicate Melnikoff et al. (2020)

original findings on the standard AMP are questions for which the current authors have no sound answers to offer at the moment. These are open questions for future empirical research.

More importantly, the present research highlights the value of a structural fit approach. When using the intentional version of the AMP, the explicit/intentional evaluative measure was sensitive to the plan status factor. Yet, this demonstration leaves us in a state of uncertainty. Specifically, it is possible that the prepared reflex's dissociation hypothesis in its *mental model assumption* is incorrect. That is, it may be incorrect to posit that deactivating a plan from working memory affects only non-attributive relations. For instance, both attributive and non-attributive relations may be deactivated, or only attributive relations exist and may be deactivated after a planned action is made inactive. Alternatively, or complementarily, it is possible that the prepared reflex's dissociation hypothesis in its *measurement assumption* is incorrect. That is, it may be incorrect to assume that explicit/intentional measures of attitudes do not capture the spread of non-attributive information in the stimulus-planned action link). Because we do not know how to design tasks that are sufficiently process-pure to capture only attributive or only non-attributive relations, the structural fit approach leaves us in a state of relative uncertainty. A Process Dissociation approach should be favored, provided a task can be designed that allows creating conditions where attributive and non-attributive relations operate in the same or opposite directions.

Conclusion

As we revisited a recent dissociation hypothesis using structurally fitted tasks, unsuspected findings emerged that questioned the validity of the proposed theory. Despite recommendations made by Payne et al. (2008) almost fifteen years ago, structural fit approaches have been rarely used and, to our knowledge, were never used to revisit a past dissociation. We hope the current adversarial collaboration highlights the value of such tests

and encourages social cognition researchers to engage in similar efforts. This may include designing new tasks that are structurally fitted counterparts of existing "implicit" measures (e.g., a structurally-fitted IAT that controls for whichever process-related feature of the standard IAT is of theoretical interest).

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievit, R. A. (2021). Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.2

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition: Basic processes* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Bengayev, E. (2020, July 27). Running Project Implicit's AMP from Qualtrics [Blog post]. Retrieved from https://minnojs.github.io/minnojs-blog/qualtrics-amp/

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*(3), 212–232. https://doi.org/10.1177/1088868320911325

De Houwer, J. (2018). Propositional Models of Evaluative Conditioning. *Social Psychological Bulletin, 13*(3), 1-21. https://doi.org/10.5964/spb.v13i3.28046

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–109). New York: Academic Press.

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science*, *14*(4), 574–595. https://doi.org/10.1177/1745691619826015

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., & Bodenhausen, G. V. (2011). The Associative–Propositional Evaluation Model. *Advances in Experimental Social Psychology, 44*, 59–127. https://doi.org/10.1016/b978-0-12-385522-0.00002-0

Gawronski, B., & Bodenhausen, G. V. (2014). Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model. *Social and Personality Psychology Compass, 8*(8), 448–462. https://doi.org/10.1111/spc3.12124

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative Conditioning From the Perspective of the Associative-Propositional Evaluation Model. *Social Psychological Bulletin, 13*(3). https://doi.org/10.5964/spb.v13i3.28024

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-Five Years of Research Using Implicit Measures. *Social Cognition, 38*(Supplement), s1–s25. https://doi.org/10.1521/soco.2020.38.supp.s1

Gollwitzer, P. M. (1999). Implementation of intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493-503. https://doi.org/10.1037/0003-066x.54.7.493

Hommel, B. (2000). The prepared reflex: Automaticity and control in stimulus-response translation. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 247-273). Cambridge, MA: MIT Press.

Hommel, B., & Wiers, R. W. (2017). Towards a unitary approach to human action control. *Trends in Cognitive Sciences, 21*(12), 940-949. https://doi.org/10.1016/j.tics.2017.09.009

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology, 27*(1), 116–159. https://doi.org/10.1080/10463283.2016.1212966

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541. https://doi.org/10.1016/0749-596x(91)90025-f

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots.* R package version 0.4.0. Retrieved from https://CRAN.R-project.org/package=ggpubr

Klauer, K. C., Dittrich, K., Scholtes, C., & Voss, A. (2015). The invariance assumption in process-dissociation models: An evaluation across three domains. *Journal of Experimental Psychology: General*, *144*(1), 198–221. https://doi.org/10.1037/xge0000044

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General*, *146*(2), 194–213. https://doi.org10.1037/xge0000239

Lange, K., Kühn, S., Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of Web servers supporting online studies. *Plos One*, *10*(7), e0134073. https://doi.org/10.1371/journal.pone.0134073

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.5.2-1. Retrieved from https://CRAN.R-project.org/package=emmeans

Mathot, S., & March, J. (2021, February 10). Conducting linguistic experiments online with OpenSesame and OSWeb. https://doi.org/10.31234/osf.io/wnryc

Mathôt, S., Schreij, D., & Theeuwes, J. (2011). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J.W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 204-217). New York: Guilford.

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Melnikoff, D. E., Lambert, R., & Bargh, J. A. (2020). Attitudes as prepared reflexes. *Journal of Experimental Social Psychology, 88,* 103950. https://doi.org/10.1016/j.jesp.2019.103950

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* R package version 0.9.12-4.2. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Payne, B. K. (2008). What Mistakes Disclose: A Process Dissociation Approach to Automatic and Controlled Processes in Social Psychology. *Social and Personality Psychology Compass, 2*(2), 1073–1092. https://doi.org/10.1111/j.1751-9004.2008.00091.x

Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *European Review of Social Psychology, 20*, 272-314. https://doi.org/10.1080/10463280903162177

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*(1), 16–31. https://doi.org/10.1037/0022-3514.94.1.16

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal*

of Experimental Social Psychology, 44(2), 386-396. https://doi.org/10.1016/j.jesp.2006.12.008

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*(9), 1043–1056. https://doi.org/10.1037/0003-066x.45.9.1043

Roediger, H. L., Weldon, M. S., Stadler, M. L., & Riegler, G. L. (1992). Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1251–1269. https://doi.org/10.1037/0278-7393.18.6.1251

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. https://doi.org/10.3758/pbr.16.2.225

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/packageafex

Torchiano, M. (2016). *Effsize - a package for efficient effect size computation.* Retrieved from http://doi.org/10.5281/zenodo.1480624

Van Dessel, P., Cone, J., Gast, A., & De Houwer, J. (2019). The impact of valenced verbal information on implicit and explicit evaluation: the role of information diagnosticity,

primacy, and memory cueing. *Cognition and Emotion*, *34*(1), 74–85. https://doi.org/10.1080/02699931.2019.1594703

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis.* New York: Springer.

Yonelinas, A. P., & Jacoby, L. L. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory & Cognition*, *40*(5), 663–680. https://doi.org/10.3758/s13421-012-0205-5

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

Zerhouni, O., Bègue, L., & O'Brien, K. S. (2019). How alcohol advertising and sponsorship works: Effects through indirect measures. *Drug and Alcohol Review*, *38*(4), 391–398. https://doi.org/10.1111/dar.12929

Zlotnick, E., Dzikiewicz, A. J., & Bar-Anan, Y. (2015). Minno.js (Version 0.3) [Computer software].