The influence of pre-training evaluative responses on approach-avoidance training outcomes

Anand Krishna & Andreas B. Eder

University of Würzburg, Germany

In press, Cognition and Emotion

Author Note

Correspondence author: Anand Krishna, Department of Psychology, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany. E-mail: krishna@psychologie.uni-wuerzburg.de; Phone: +49 931 31 86674, Fax: +49 931 31 82812, ORCid: https://orcid.org/0000-0003-3838-9210.

Andreas B. Eder, Department of Psychology, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany. E-mail: andreas.eder@uni-wuerzburg.de; Phone: +49 931 31 83336, Fax: +49 931 31 82812.

Abstract

Approach-avoidance training (AAT) has been shown to be effective in both clinical and laboratory research. However, some studies have failed to show effects of AAT. Therefore, finding moderators of the AAT effect is a priority for further research. We investigate the moderating effect of pre-training evaluative responses towards familiar AAT targets. In particular, we test predictions: (a) that congruent responses (i.e., approach to positive targets and avoidance of negative targets) increase liking, whereas incongruent responses decrease liking; (b) that training is more effective when it can strengthen existing positivity or negativity; and (c) that ambivalence increases AAT effects. Two experiments (total *N* = 132) implemented an AAT with local soft-drink brands after measuring initial positive/negative explicit evaluative components and implicit liking towards the brands. Results show no reliable evidence for training effects on consumption or rating of drinks, but participants showed more implicit liking of approached drinks than avoided drinks. Furthermore, the magnitude of implicit liking measured pre-training was positively related to the size of the training effect. Ambivalence had no direct effect on the training outcomes. These results partially support the congruency prediction and underline the importance of implicit liking prior to AAT as a moderator for AAT effects.

Keywords: approach-avoidance training; attitudes; preference; consumption; pre-existing preferences

The influence of pre-training evaluative responses on approach-avoidance training outcomes

Human beings are motivated to approach pleasurable endstates and avoid painful ones (Eder, Elliot, & Harmon-Jones, 2013). This tendency can even cause automatic impulses that are opposed to an individual's explicit goals, such as eating sugary food even though attempting to restrict intake (Hofmann, Friese, & Roefs, 2009). Consequently, several studies have investigated whether automatic approach-avoidance tendencies can be changed with so-called 'approach-avoidance training' (AAT) procedures (e.g. Kawakami, Phills, Steele, & Dovidio, 2007). Participants repeatedly perform approach-related actions (e.g. pulling a joystick lever) in response to a specific class of stimuli (e.g., soft drinks) and avoidance-related actions (e.g. pushing a joystick lever) in response to another (e.g., alcohol-related stimuli). The typical finding is that the approach-trained stimuli are preferred relative to the avoidance-trained stimuli in behavioural and liking measures. For example, one of the first AAT studies to apply this procedure showed that AAT reduced implicit bias towards a stereotyped group (Kawakami et al., 2007), whereas other studies have extended the effects to reducing fear of spiders (Jones, Vilensky, Vasey, & Fazio, 2013) or mitigating contamination anxiety (Amir, Kuckertz, & Najmi, 2013). One of the most prolific fields of AAT application has been that of consumption behaviour. Wiers and colleagues (2011) trained alcoholic patients to avoid alcohol-related stimuli and found that their likelihood of relapse was reduced following the training. This striking finding has been replicated in a large clinical sample (Eberl et al., 2013), although similar results were not obtained with subclinical samples of social or at-risk drinkers (Lindgren et al., 2015) or via online interventions (Wiers et al., 2015). Similarly, several other studies in the fields of food consumption (Becker, Jostmann, Wiers, & Holland, 2015) and fear reduction training (Krypotos, Arnaudova, Effting, Kindt, & Beckers, 2015; van Uijen, van den Hout, & Engelhard, 2015) have also failed to produce AAT effects. AAT has also been investigated with neutral, unfamiliar targets. In this research, evaluative responses to nonwords (Van Dessel, Gawronski, Smith, & De Houwer, 2017), fictitious social groups (Van Dessel, Eder, & Hughes, 2018), or fictitious soft drink brands (Zogmaister, Perugini, & Richetin, 2016) were changed by AAT procedures. While these studies were successful in producing an AAT effect, other studies failed to find an effect, some of them even providing Bayesian evidence for the null hypothesis when using unfamiliar stimuli (Krishna & Eder, in press; Van Dessel, De Houwer, & Gast, 2016; Van Dessel, De Houwer, Roets, & Gast, 2016; Vandenbosch & De Houwer, 2011).

These mixed findings on the effectiveness of AAT point towards the existence of unknown moderating variables. Zogmaister and colleagues (2016) identified motivational states as determinants of AAT's effectiveness, finding that participants liked approached (relative to non-approached) soft drinks more in a state of thirst and approached potato chips more when hungry. The authors argue that the thirst state may improve evaluations of the target drinks before training, which in turn may interact with the training of approach responses. Thus, evaluations of the training stimuli that exist before the training could be important moderators of the effectiveness of AAT procedures. This was the main research question of the present study.

Effects of pre-training stimulus valence on AAT outcomes: A few hypotheses

Effects of AAT procedures on the liking of target stimuli have most typically been explained with the learning of a connection between a specific valence and the target. Cognitive theories have postulated that this valence comes from the motivational compatibility between positive/negative targets and approach/avoidance actions (Centerbar & Clore, 2006), or the valence of the associated action goal (affective response-coding; Eder & Rothermund, 2008), or from a propositional inference of the likability of approached and avoided stimuli (Van

Dessel et al., 2017; Van Dessel, Hughes, & De Houwer, in press).

The motivational compatibility account assumes that reacting with approach to positive stimuli and reacting with avoidance to negative stimuli are congruent motivational responses that enhance liking of the stimuli (Centerbar & Clore, 2006). This *congruency hypothesis* predicts that typical AAT effects should occur for positive targets, but reversed AAT effects should occur for negative targets: avoiding negative targets should improve liking, whereas approaching them should reduce liking. A similar prediction could be derived from the response-coding account (Eder & Rothermund, 2008). According to this account, both approach to positive stimuli and (successful) avoidance of negative stimuli should result in a pleasurable end-state, which in turn should improve the liking of the successfully avoided stimuli (Eder & Dignath, 2014; see also Mertens, Van Dessel, & De Houwer, 2018). It should be noted that one study seemingly rejects this hypothesis because it observed evaluative changes in an implicit measure only after training with neutral faces but not with training of angry and smiling faces (Woud, Becker, Lange, & Rinck, 2013). However, the difference between these stimulus conditions was not significant according to a conventional criterion ($p < .05$), and motivational responses to facial expressions are complex (Paulus & Wentura, 2016; Wilkowski & Meier, 2010). Therefore, it seems premature to reject a congruency hypothesis on the basis of this single study.

[Insert Figure 1 about here]

[Insert Figure 2 about here]The propositional inference account assumes that a proposition exists that positive stimuli are typically approached and negative ones typically avoided. Therefore, participants may infer from the training that they like the approached stimuli more than the avoided ones. The likelihood that a propositional inference is accepted as valid is based on its consistency (Gawronski & Bodenhausen, 2006; Van Dessel et al., in press). Hence, if participants have pre-training attitudes towards targets that are opposed to the inference they would draw from their behaviour, this inference should be less likely to influence their attitudes (e.g. the proposition "I dislike spiders" is more consistent with the proposition "I avoided spiders, therefore I do not like them" than with the proposition "I approached spiders, therefore I like them"). This account hence suggests a *strengthening hypothesis*: approach training should have stronger effects for positive targets, and avoidance training should have stronger effects for negative targets. A comparison of the congruency and strengthening hypotheses separated by positive and negative attitude components is shown in Figure 1. Both hypotheses make similar predictions for the role of positive attitude components (albeit different in expected magnitude), but for negative attitude components, the predictions differ, especially for the avoided target. Importantly, strengthening of one motivational response (e.g., approach) is offset by the relative weakening of the other response (e.g., avoidance), which means that specific strengthening effects can only be detected if changes in positive and negative attitude components are separately assessed. If both of these components are assessed by a single measure, the prediction for the strengthening hypothesis simplifies to a main effect of AAT and a main effect of attitudes with no interaction (see Figure 2). In comparison to the congruency hypothesis, the critical difference is therefore found in the avoided target, for which the predicted slope differs.

Another theoretical construct of interest is attitudinal ambivalence. Attitudes are ambivalent when they contain both strong positive and strong negative components simultaneously (Thompson, Zanna, & Griffin, 1995), as opposed to being univalent (*either* strong positive *or* strong negative) or indifferent (*neither* strong positive *nor* strong negative; Kaplan, 1972). In alcohol AAT studies, it is likely that the heavy drinkers had ambivalent attitudes about alcohol consumption (for evidence see Stritzke, McEvoy, Wheat, Dyer, & French, 2007). Wiers and colleagues (2011) proposed that AAT can change the *relative* accessibility or weight of positive or negative components, making the attitude less ambivalent and more univalent. Therefore, the *ambivalence hypothesis* states that AAT effects should be enhanced when there is more attitudinal ambivalence before the training (see Figure 3).

[Insert Figure 3 about here]

A previous study of Jones and colleagues (2013) trained participants to approach spiders and insects—target stimuli deemed unambiguously negative. They found that AAT was effective even for these targets, which would seem to contradict the ambivalence hypothesis. However, it is important to note several points: First, Jones et al. did not measure attitudinal ambivalence. It is therefore possible that their AAT effects were driven by individuals who had some positive evaluative components towards spiders (Ellwart, Rinck, & Becker, 2006). In addition, the study presented targets subliminally, which may not have the same effect in biasing subsequent disambiguation as consciously processed stimuli (see Van Dessel et al., 2016). Finally, it is possible that ambivalence is a relevant moderator, but does not completely determine AAT effects.

The present research tested the congruency, strengthening and ambivalence hypotheses by retraining motivational action tendencies towards two familiar, sugary soft drinks. People often have mixed positive and negative attitudes towards the consumption of sugary beverages (Kassem, Lee, Modeste, & Johnston, 2003). Therefore, we measured participants' initial positive and negative evaluative components as well as their implicit liking of the soft drink brands before the AAT task. We then measured participants' consumption of these soft drinks as well as their implicit and explicit evaluations of the drinks after training. It should be noted that past studies have found effects of AAT both on implicit evaluative measures (e.g. Mertens et al., 2018) and on explicit ratings (e.g. Centerbar & Clore, 2006), so we wanted to investigate our hypotheses with regard to both implicit and explicit pre-training (initial) evaluative responses. However, we are not aware of any validated implicit ambivalence measure, so ambivalence was assessed using only explicit ratings.

Experiment 1

Method

Participants

Data collection was planned with $N = 71$ participants to detect an AAT effect with a minimum size of $d_z = .3$ with a statistical power of 1-beta = .80 and the alpha-level set to .05. Due to a technical issue, data from the first 20 participants were incomplete and were replaced by additional participants. The final sample had 75 participants (63 female, $M_{age} = 24.9$ years, $SD = 5.5$). The study procedure was approved by the ethics committee of the Department of Psychology, University of Würzburg (GZEK 2013-14).

Apparatus and Stimuli

A Logitech Attack 3 Joystick was attached to the computer. Training stimuli were pictures of commercially available cans of Schwip Schwap© and Mezzo Mix© (240 x 400 pixels). These soft drinks have a similar taste. Pictures were taken of upright cans. These were tilted to the left or right for the behavioural assessment task. For the consumption test, cooled cans of each drink were provided (0.33l each). Target words for the affective priming task are listed in the supplementary materials.

Procedure

After giving informed consent, participants completed the implicit and explicit measures of pre-training evaluative responses (order counterbalanced). After a filler task of roughly 20 minutes, participants completed the AAT task. Thereafter, they tasted the soft drinks (the consumption measure). Next, they completed the post-training attitude assessment consisting of an explicit rating of the drinks and an affective priming task (counterbalanced order).

Explicit pre-rating. For a separate assessment of positive and negative evaluative components (additionally indexing attitudinal ambivalence), participants rated the drinks' positive and negative qualities separately (Kaplan, 1972). They rated one soft drink on how positive it was *ignoring* any negative qualities, on a scale of 1 (not at all) to 4 (very positive). They then rated how certain they were about this judgment on a Likert scale of 1 (hardly certain) to 9 (very certain) as an index of prior experience with the brand. Next, they rated how negative the same soft drink was, this time ignoring any positive qualities, on an analogous 4-point scale before again indicating how certain they were about this judgment. Thereafter, they performed these ratings for the other soft drink. The order of the soft drinks was randomized.

Pre-training implicit liking. Pre-training implicit attitudes towards the soft drinks were assessed using an affective priming procedure (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). Participants responded to target words by categorizing them as positive or negative using the "A" and "L" keys on the keyboard. The mapping of the response keys was counterbalanced across participants. Task instructions were to categorize the words as fast and as accurately as possible. Participants were also told that before each word, a picture would appear, but that this picture was irrelevant for the task at hand and could therefore be ignored. Each trial started with a fixation cross for 750 ms, followed by a soft drink picture for 200 ms, a blank screen for 50ms, then the target word appeared. An error message appeared for 2,000 ms if the response was incorrect or slower than 2,000 ms. After 500 ms, the next trial was initiated. Participants completed two blocks of 48 trials each, with 12 positive target words and 12 negative target words per drink.

AAT task. After a set of unrelated filler reaction time tasks, participants were offered as many salty snacks as they liked, after which they rated their feeling of thirst on a 7-point Likert scale from 1 (not at all thirsty) to 7 (very thirsty), then their current mood on another 7-point Likert scale from -3 (very bad) to +3 (very good). Participants were instructed to respond to pictures of cans of Mezzo Mix© and Schwip Schwap© by pulling and pushing a joystick lever based on the brand. The mapping of responses to brands was counterbalanced across participants. In each trial, participants saw an asterisk in the middle of the screen as a fixation point for 200ms, followed by a blank screen for 100ms, and a picture of a soft drink. If participants responded correctly, the soft drink picture on the computer screen was zoomed towards (if pulled correctly) or away (if pushed correctly) from the participant. This zooming effect lasted 300ms. An error message appeared for 2,000ms if the response was incorrect or slower than 2,000ms. The next trial began after 500ms. If the joystick lever was not centred at start, an error message appeared for 2,000ms, after which the trial restarted. Participants performed 10 blocks with 10 training trials for each soft drink (200 trials total). After 100 AAT trials, a behavioural assessment task was intermixed to check the effectiveness of our AAT procedure for a change in approach-avoidance tendencies. This assessment task was identical to the AAT task except that the instruction was now to respond to the orientations of the soft drink cans that were now slightly tilted to the left or to the right. The mapping of the lever movements to the left-right orientation was counterbalanced across participants. Half of the drink pictures afforded a response that was congruent with the trained response during the AAT task, while the other half of the pictures afforded a response that was different from the trained response. Participants completed two blocks of 20 trials each of the behavioural assessment task before returning to the second half of the AAT trials. It was expected that congruent lever responses would be performed faster and with less errors.

Consumption test. Next, the participant was asked to drink as much Schwip Schwap© and Mezzo Mix© as they wanted. The experimenter offered one can of each drink. If participants finished either can, the experimenter brought more until the participant was satisfied. After the participant had finished tasting the soft drinks, the experimenter unobtrusively weighed (in grams) how much the participant had drunk of each.

Post-training attitude assessment. After the consumption test, participants completed a second affective priming task (identical to the first) and provided a global evaluation of each drink on a 9-point Likert scale from -4 (very negative) to +4 (very positive), each time followed by a rating of how certain they were of this judgment on a 9-point Likert scale similar to the explicit pre-rating. The order of the implicit and explicit tasks was counterbalanced across participants. Finally, participants were asked about their suspicions on the purpose of the study and completed measures of state reactance (perceived threat to choice and anger subscales retrieved from "Consumer Health Informatics Research Resource - Age," n.d.) and trait reactance (Hong & Page, 1989).

Results

Data preparation

For all tasks, timeouts were coded as errors. Participants were excluded from analysis based on error rates in the AAT task (excluded from all analyses) and in the pre- and post-training affective priming tasks (excluded from analyses including indices derived from these tasks). Errors and individual RT outlier trials were eliminated before calculating mean RTs in the affective priming tasks. For details on error-based participant and individualized trial outlier elimination procedures, see the supplementary materials. Two participants were excluded due to excessive AAT errors (remaining sample error rate: M = 2.7%, SD = 1.7, range 0-8%). One additional participant was eliminated due to errors in the affective priming tasks. For the behavioural assessment task, the first two trials were removed as practice trials. For the affective priming tasks, the first four trials were removed as practice trials. After exclusion of outlier and error trials, mean RTs to positive targets following a specific drink were subtracted from mean RTs to negative targets for the same drink to form an index of implicit liking for that drink (with higher values indicating more liking). An analogous score was computed for the error rates by subtracting errors to positive targets from errors to negative targets.

Explicit ambivalence scores for each drink were calculated for explicit positive (P) and negative (N) evaluative components according to the Griffin formula (Thompson et al., 1995). This formula accounts for both intensity and similarity of these components by subtracting their absolute difference from their mean (i.e., (P+N)/2 – |P-N|), with higher scores indicating higher ambivalence. Certainty scores for each evaluative component were averaged to form an index of the certainty of the pre-training evaluation.

Raw data, data preparation and analysis scripts for all experiments may be found on the Open Science Framework (osf.io/t7kmf).

Descriptive data

Participants indicated moderately positive mood (*M* = 1.1, *SD* = 1.1) and moderate thirst (*M* = 4.4, *SD* = 1.1).

Tables

Table 1 shows descriptive data for pre-training evaluative measures towards both drinks. Participants found both drinks moderately positive and moderately negative and were moderately certain about their ratings, implying some degree of previous experience with both soft drinks. The evaluations of the soft drinks did not significantly differ on any dimension.

[Insert Table 1 about here]

Analyses and model specifications

The consumption, explicit ratings, and post-training affective priming measures were analysed for effects of pre-training evaluations using restricted maximum-likelihood linear mixed models. All continuous variables were grand-mean centred to increase interpretability. Subject intercept was entered as a random factor in all models and subject-level means for all continuous predictors were entered as covariates in all models to control for subject-level response biases (except for the explicit ambivalence model, as the calculation of ambivalence scores renders the subject-level mean meaningless as a control for response biases). For simplicity, subject-level mean effects are not reported below (see the supplementary materials for a complete report of the models).

Both the congruence and the strengthening hypotheses were tested for explicit measures using a model including positive and negative evaluative components as fixed effects, AAT (approach, avoidance) as a dummy variable and the two-way interactions between AAT and the positive or negative evaluative component. This model was used to estimate post-training drink consumption and post-training explicit rating. The critical statistical test for these models was that of the two-way interaction terms, with both the congruency and strengthening accounts predicting significant interactions, albeit with different patterns: the congruency account predicted stronger AAT effects with increasing positive evaluative components, but weaker AAT effects with increasing negative evaluative components, whereas the strengthening hypothesis predicted a stronger positive relationship between positive evaluative components and post-training liking for the approached target and a stronger negative relationship between negative evaluative components and post-training liking for the avoided target (see Figure 1).

An analogous model was calculated for the pre-training implicit measure that included the fixed effects implicit liking (RTs) and dummy-coded AAT as well as their interaction term. This model was used to estimate post-training drink consumption and post-training implicit liking. Due to the relative nature of the implicit liking measure, the strengthening hypothesis predicted a significant AAT effect and a significant main effect of pre-training implicit liking, but the congruence hypothesis predicted a significant two-way interaction term such that AAT effects increase with implicit liking (see Figure 2). Similar models were calculated for the implicit liking indices based on error rates where appropriate, but they produced no effect in any analysis and are therefore not reported.

To test the ambivalence hypothesis, a model was calculated that included explicit ambivalence and dummy-coded AAT as well as their interaction term. This model was used to estimate effects of AAT on post-training drink consumption, post-training implicit liking and post-training explicit rating. The critical term was the two-way interaction, which was predicted to show greater AAT effects with increasing explicit ambivalence (see Figure 3).

Moderator analyses for thirst (cf. Zogmaister et al., 2016) and both state and trait reactance over both experiments and discussions of the results may be found in the supplementary materials.

Drink consumption

The explicit evaluative components model revealed an effect of positive evaluative components that was close to the standard criterion of statistical significance, $F(1,68.2) = 3.82$, $B = 18.3$, $p = .055$. This result suggests that participants with more positive evaluative components towards a drink consumed more of it, while negative evaluative components had no effect on consumption. No other effects in this model were significant (all *ps* > .5). The implicit liking model revealed a tendency for an interaction between pre-training implicit liking and AAT, $F(1,87.8) = 3.34$, $p = .070$. No other effects in this model were significant (all *ps* > .2). As displayed in Figure 4, the slope for pre-training implicit liking was negative for avoided drinks, $B = -.27$, $p = .052$, and near zero for approached drinks, $B = .03$, $p = .812$. This result is partially consistent with the congruence hypothesis – training effects are reversed when implicit liking is low, but this seems to be driven only by the avoidance training with no corresponding effect for the approach training. It should be noted that the interaction effect weakens ($p > .1$) when outlier trials and excluded participants are included in the analysis. The explicit ambivalence model produced no significant effects (all *ps* > .7).

[Insert Figure 4 about here]

Implicit preference

The implicit liking model produced a significant main effect of AAT, $F(1,68.1) = 6.26$, $p = .015$, which was qualified by a significant interaction with pre-training implicit liking, $F(1,102.9) = 4.07$, $p = .046$. Similarly to the model estimating consumption, the slope of pre-training implicit liking was negative for avoided drinks, $B = -.29$, $p = .013$, and near zero for approached drinks, $B = -.02$, $p = .869$ (see Figure 5). This result is partially consistent with the congruency hypothesis, as avoidance training is more effective when the target's pre-training valence is relatively positive, but there is no corresponding effect for approach training. The explicit ambivalence model showed an analogous main effect of AAT on post-training, but no other effects (all $p > .2$).

[Insert Figure 5 about here]

Explicit rating

The explicit evaluative components model failed to achieve convergence when the random intercept was included, indicating that the variance associated with the random intercept was insufficient to achieve a good model fit. The random intercept was therefore removed for this analysis. The model produced a main effect of positive evaluative components, $F(1,138) = 4.54$, $B = .61$, $p = .035$, and no other significant effects (all $ps > .4$). Drinks that were rated positively before the training were also rated more positively after the training, whereas pre-training negative components had no effect. The explicit ambivalence model produced no significant effects (all $ps > .4$).

Behavioural assessment task

Performance in the behavioural assessment task was analysed using one-tailed paired-samples t-tests. There was no significant effect of congruence on either the RT measure, $t < 1$, or the error measure, $t(72) = 1.53$, $p = .130$.

Discussion

We obtained an effect of AAT in our study, though only on the implicit liking measure and not in the explicit ratings or consumption measures. More importantly for the present discussion, the post-training implicit evaluation measure and (less robustly) drink consumption were affected by AAT differently depending on participants' pre-training implicit evaluations towards the drinks. Consistent with the congruency hypothesis, the effectiveness of the training procedure depended on pre-training implicit liking towards the drinks.

Experiment 1 provided no support for any effect of evaluative ambivalence on AAT. However, a potential weakness of the measurement procedure was that participants had to list their evaluative components sequentially. Such a measurement might drive participants to resolve their ambivalence directly in order to answer the individual items, leaving little space for ambivalence-based AAT effects. Although this procedure has often been used to measure ambivalent attitudes in the past (e.g., Kaplan, 1972), a simultaneous assessment of both evaluative components might be better suited. Furthermore, the behavioural assessment task provided no evidence that our AAT procedure had an effect on automatic approach-avoidance tendencies towards the drinks. We believe that processing of the left-right orientation was too shallow for a processing of the irrelevant feature (brand) on the reaction times (Spruyt, De Houwer, & Hermans, 2009). Therefore, Experiment 2 sought to replicate Experiment 1 while implementing a more sensitive measure of ambivalence and an improved version of the behavioural assessment task.

Experiment 2

Method

Participants

As power analyses cannot be readily calculated for mixed linear models, we set a desired sample size of 60 for Experiment 2 based on economic considerations. Sixty volunteers from the Würzburg area (44 female, M_{age} = 27.5, *SD* = 10.2) took part.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1 in all ways with the following exceptions:

Explicit pre-rating. An evaluative space grid was now used for a simultaneous measurement of positive and negative evaluative components (Larsen, Norris, McGraw, Hawkley, & Cacioppo, 2009). Participants saw a 5x5 grid with anchors from "not at all" to "extremely". On the vertical axis, higher values indicated higher negativity, while on the horizontal axis, higher values indicated higher positivity. Thus, the bottom left square corresponded to "not at all positive/not at all negative", the top left to "not at all positive/extremely negative", the bottom right to "extremely positive/not at all negative", and the top right to "extremely positive/extremely negative." After selecting a position on the grid and confirming their selection, participants were asked to rate how certain they felt about their judgment on a 9-point Likert scale from 1 (hardly certain) to 9 (very certain). The drinks were evaluated in random order.

Behavioural assessment task. The behavioural assessment task was identical to that of Experiment 1 except that the relevant feature for the selection of a lever response was now whether the can was presented normally or inverted (flipped upside-down). Thus, participants were forced to focus on the brand logo in order to see whether it was inverted, which should increase its influence on behavioural performance.

Results

Data preparation.

Outlier identification and data preparation was the same as in Experiment 1. One participant was eliminated due to a high error rate in the AAT task (remaining sample error rate: M = 3.2%, SD = 1.9, range 0-8%). One participant was excluded from analyses because of a high error rate in the pre-training affective priming task. Details of exclusion thresholds are provided in the supplementary materials.

Descriptive data

Participants indicated moderately positive mood (*M* = 1.0, *SD* = 1.3) and moderate thirst (*M* = 5.1, *SD* = 1.2).

Table 2 shows descriptive data for pre-training evaluation measures towards both drinks. As for Experiment 1, participants evaluated both drinks moderately positively and moderately negatively, and they were moderately certain about their ratings, implying some degree of previous experience with the target drinks. No ratings differed between drinks (all *ps* > .30). Differences in absolute values to Experiment 1 on the positive/negative components and ambivalence scores are due to the inclusion of a zero point in the evaluative grid in this experiment.

[Insert Table 2 about here]

Analyses and model specifications

Model specifications and applied analyses were the same as in Experiment 1. The error-based index of pre-training implicit liking produced no significant effects in any analysis and is not reported below.

Drink consumption

The explicit evaluative components model showed a tendency for an interaction between positive evaluative components and AAT, $F(1,56.0) = 3.59$, $p = .063$, and no other significant effects (all *ps* > .2). The slope for positive evaluative components was near zero for the avoided drink, $B = .28$, $p = .979$, but larger for the approached drink, $B = 15.72$, $p = .163$. Participants consumed more of a drink they had previously seen as positive after approaching it, but not after avoiding it. This result is partially consistent with both the congruency and the strengthening hypotheses, but cannot differentiate between the two without a corresponding effect on negative evaluative components. The implicit liking model produced no effects (all *ps* > .3), nor did the explicit ambivalence model (all *ps* > .6).

Implicit preference

The implicit liking model produced a significant main effect of AAT, $F(1,55.6) = 7.32$, $p = .009$. The interaction effect from Experiment 1 failed to achieve significance, $F(1,99.8) = 2.49$, $p = .118$ (all other *ps* > .1). However, the trend was similar to that observed before: the slope of pre-training implicit liking was smaller for avoided drinks, $B = .05$, $p = .678$, than for approached drinks, $B = .28$, $p = .027$, again providing partial support for the congruency hypothesis. The explicit ambivalence model showed no effects beyond the previously demonstrated main effect of AAT in the former (all *ps* > .3).

Explicit rating

The explicit evaluative components model produced a main effect of negative evaluative components, $F(1,53.7) = 7.67$, average $B = -.74$, $p = .008$. No other effects were significant (all *ps* > .2). The explicit ambivalence model produced no significant effects (all *ps* > .6).

Behavioural assessment task

The data from the behavioural assessment task were analysed using one-tailed within-subjects t-tests. In line with the AAT, participants were slower in incongruent trials than congruent ones, $t(58) = 3.28$, $p < .001$, $d_z = .43$, and also made more errors in incongruent trials, $t(58) = 3.65$, $p < .001$, $d_z = .48$.

Discussion

Experiment 2 partially replicated the results of Experiment 1: AAT again affected post-training implicit evaluation measures, but not explicit judgments or consumption. The previously found interaction between AAT and pre-training implicit liking was not replicated for consumption and did not achieve significance for the implicit evaluation measure.

However, Experiment 2 produced a tendency for positive explicit evaluative components to influence consumption only for approached drinks. Once again, there were no effects of ambivalence scores.

Taken together, the results of both experiments indicate that pre-training evaluations may play a role for the effectiveness of AAT procedures, especially pre-training implicit evaluations. However, it is unclear whether these effects are reliable. Therefore, the data of both experiments were pooled into a mega-analysis in order to identify robust effects.

Mega-Analysis

Results

Data preparation

Data was prepared for each experiment individually as described in their respective results sections. As the explicit evaluative component scales and ambivalence measure varied in scaling between the experiments, these scores were z-standardized for each experiment prior to analysis.

Drink consumption

The explicit evaluative components model produced a significant main effect of positive evaluative components, $F(1,126.1) = 4.16$, average $B = 11.79$, $p = .043$, and no other significant effects (all *ps* > .1). Neither the implicit liking model (all *ps* > .3) nor the explicit ambivalence model (all *ps* > .7) produced any effects.

Implicit preference

The implicit liking model produced a significant main effect of AAT, $F(1,126.4) = 10.95$, $p = .001$. The interaction effect of AAT and pre-training implicit liking was significant, *F*(1,208.6) = 5.89, *p* = .016. The slope of pre-training implicit liking was smaller for avoided drinks, *B* = -.13, *p* = .126, than for approached drinks, *B* = .11, *p* = .212 (see Figure 6). These results support the congruency hypothesis, although there is no reversal of training effects for drinks that were evaluated negatively before the training. However, it is important to note that the interaction effect no longer achieves significance if a Bonferroni correction for multiple testing of the congruency hypothesis is applied (α = .05/6 = .008)[1]. All other effects were not significant (with *ps* > .80). The explicit ambivalence model showed no effects beyond the established main effect of AAT in the RT (all *ps* > .4) and error measure (all *ps* > .1).

[Insert Figure 6 about here]

Explicit rating

The explicit evaluative components model produced a significant main effect of positive evaluative components, *F*(1,128.1) = 5.75, average *B* = .55, *p* = .018, such that participants who indicated stronger pre-training positive evaluative components towards a drink liked that drink more in the post-rating. No other effects were significant (all *ps* > .2). The explicit ambivalence model produced no significant effects (all *ps* > .4).

General Discussion

Two experiments examined the role of initial explicit and implicit evaluations on AAT effects for soft drinks. Before training, participants rated positive and negative aspects of two popular local soft drink brands and completed an affective priming task with these soft drink

[1] The correction reflects a total of six tests: interaction terms between AAT and two explicit evaluative components as well as one implicit liking measure, each tested over two dependent variables.

brands presented as primes. Then, they were trained to approach one of the brands and avoid the other in a joystick AAT task with zooming effects. After the training, drink consumption and liking for the drinks were measured using explicit ratings and an affective priming task. The results of the mega-analysis showed no AAT effects on consumption or explicit liking, but there was an AAT effect on implicit liking as well as an interaction of AAT with the pre-training implicit evaluation measure: the AAT effect increased with pre-training implicit liking towards the drinks. There was no support for any interaction between pre-training explicit evaluative component ratings and AAT or pre-training explicit ambivalence scores and AAT.

Results provided no unqualified support for the ambivalence or strengthening hypotheses (but see supplementary materials for exploratory moderator analyses). Instead, they are consistent with the congruency hypothesis: a standard AAT effect on implicit liking (i.e., more liking of the approached drink relative to the avoided drink) occurred when the pre-training implicit liking towards the targets was relatively strong. In contrast, no AAT effect was observed when the drinks were evaluated negatively before the training. This finding must be interpreted with some caution, however, as the statistical evidence for it is not robust to correction for multiple tests. It also is unclear why the AAT effect was not reversed for negative drinks as predicted by the congruency account. However, it is possible that implicit attitudes towards sugary soft drinks were not negative enough to induce a reversal. Future studies might increase the magnitude of the effect by contrasting targets with clear negative pre-training attitudes directly against those with clear positive pre-training attitudes.

The results provide tentative support for both motivational compatibility accounts (Centerbar & Clore, 2006) and response-coding accounts (Eder & Dignath, 2014; Eder & Hommel, 2013; see also Mertens et al., 2018) of AAT. The former assume that performed actions that are motivationally compatible with their targets impact attitudes towards the targets positively, while the latter assume that actions leading to a positive outcome do the same. Critically, both therefore predict standard AAT effects only for positive targets, as *avoiding* negative targets is motivationally *compatible* and leads to a relatively *positive* outcome. We extend previous findings by Centerbar and Clore (2006) by demonstrating an effect consistent with motivational congruency in a larger sample size with familiar (rather than unfamiliar) target stimuli, showing that such congruency effects may also occur when participants have significant prior knowledge of the targets. The present results suggest that AAT effects are strongest for positive targets, which is of practical relevance for applied research on AAT as an intervention. Furthermore, future AAT research may want to focus more on negative target stimuli to investigate what factors can account for the discrepant findings between the present study and previous studies that obtained typical AAT effects with ostensibly univalent negative stimuli (Amir et al., 2013; Jones et al., 2013).

The results are not in line with a propositional inferential account, which assumes that propositions must be compatible with other activated propositions for an inference (Van Dessel et al., in press). It is unclear how this account could explain how the proposition "I dislike this drink" (based on the negative pre-training implicit liking) and the proposition "I avoided this drink" (based on the training procedure) should result in the inference "I like this drink" (the observed response indexed by the implicit liking measure). Though this trend towards increased liking was not statistically robust in itself, explaining why it is significantly different to the effect for liked drinks is difficult for the propositional account without additional assumptions. For example, it is possible that the implicit liking measure did not reflect precisely those propositions. More specifically, the pre-training implicit liking measure may instead have reflected "This drink is unhealthy", leading to the post-training inference "avoiding this unhealthy drink was good", which in turn may have influenced the post-training measure. Furthermore, the fact that the congruence effect was only found on the implicit liking measure may indicate that this measure captured inferences that were based on well-practiced rules and therefore fast and easily activated; consequently, the absence of an effect on the explicit and consumption measures may reflect increased interference by more complex inference rules (Van Dessel et al., in press) or more elaborate validity testing of inferences (Gawronski & Bodenhausen, 2006). Therefore, the present studies' findings do not challenge the propositional inference account when auxiliary assumptions are made. However, they underline the necessity for this account to specify in more detail what inferences are expected under which circumstances.

This study's findings may seem to be at odds with previous studies that obtained AAT effects with ostensibly neutral stimuli (e.g., Huijding et al., 2009; Van Dessel et al., 2018). It should be noted, however, that the absence of a clear attitude before training does not imply that an AAT effect is not possible. Effects might only be weaker under such circumstances. Furthermore, it is questionable whether the ostensibly neutral stimuli in previous studies were really neutral in an absolute sense (Perugini, Richetin, & Zogmaister, 2012). Studies have shown that even ostensibly "neutral" stimuli, such as nonwords (Bakhtiari, Körner, & Topolinski, 2016) or Chinese ideographs (Centerbar & Clore, 2006), can possess a subtle valence. In addition, most studies with neutral stimuli did not check the neutrality before training. Therefore, it is possible that these stimuli possessed a subtle valence that in turn influenced the effectiveness of the AAT procedure.

One limitation of the present study is that we could not differentiate between pre-training evaluative effects for novel stimuli against familiar stimuli. Thus, it is possible that the congruence effects observed in the present research hold only for stimuli which are familiar to participants (but see Centerbar & Clore, 2006). Future studies should examine this limitation with an experimental manipulation of pre-training valence.

A further limitation of our study is the lack of any consistent effect on explicit rating or consumption measures. In general, it must be noted that AAT research is inconsistent with regard to which measure best shows the effect. For instance, Wiers and colleagues (2010) observed an effect of AAT procedures on implicit but not on explicit evaluations, while Huijding and colleagues (2009) observed an AAT effect on explicit but not on implicit evaluations. Assessing AAT outcome measures on behavioural, implicit, and explicit measurement levels increases the overall sensitivity of the test (see Breckler, 1984), but it also increases the false positive rate and the chance for finding spurious dissociations. In our studies, both the effect of AAT and its moderation by pre-training evaluations were only found on the implicit liking measure. It is possible that processes that are less likely to impact this measure (e.g. due to requiring a longer processing period) may obscure this effect in non-speeded evaluations or consumption decisions. This possibility is bolstered by exploratory analyses that indicated that reactance and thirst moderated AAT effects for explicit, but not implicit evaluations (see supplementary materials). Clearly, more research is needed on what type of outcome measure is most appropriate for tests of AAT effects.

In conclusion, these experiments show that pre-training evaluations influence the effectiveness of AAT procedures. Future research should therefore take this into account both in stimulus selection and in identifying circumstances conducive to strong AAT effects.

Acknowledgements

We thank the research assistants who organized the experimental sessions and collected the data.

Funding Details

This work was supported by the German Research Foundation (DFG) under grant ED 201/3-1 to AE. The DFG had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

Declaration of Interest Statement

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest (such as personal or professional relationships) in the subject matter or materials discussed in this manuscript.

References

Amir, N., Kuckertz, J. M., & Najmi, S. (2013). The Effect of Modifying Automatic Action Tendencies on Overt Avoidance Behaviors. *Emotion*, *13*(3), 478–484. https://doi.org/10.1037/a0030443

Bakhtiari, G., Körner, A., & Topolinski, S. (2016). The role of fluency in preferences for inward over outward words. *Acta Psychologica*, *171*, 110–117. https://doi.org/10.1016/j.actpsy.2016.10.006

Becker, D., Jostmann, N. B., Wiers, R. W., & Holland, R. W. (2015). Approach avoidance training in the eating domain: Testing the effectiveness across three single session studies. *Appetite*, *85*, 58–65. https://doi.org/10.1016/j.appet.2014.11.017

Centerbar, D. B., & Clore, G. L. (2006). Do Approach-Avoidance Actions Create Attitudes? *Psychological Science*, *17*(1), 22–29. https://doi.org/10.1111/j.1467-9280.2005.01660.x

Consumer Health Informatics Research Resource - Age. (n.d.). Retrieved March 20, 2018, from https://chirr.nlm.nih.gov/state-reactance.php

Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best? *Developmental Cognitive Neuroscience*, *4*, 38–51. https://doi.org/10.1016/j.dcn.2012.11.002

Eder, A. B., & Dignath, D. (2014). I like to get nothing: Implicit and explicit evaluation of avoided negative outcomes. *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*(1), 55–62. https://doi.org/10.1037/xan0000005

Eder, A. B., Elliot, A. J., & Harmon-Jones, E. (2013). Approach and Avoidance Motivation: Issues and Advances. *Emotion Review*, *5*(3), 227–229. https://doi.org/10.1177/1754073913477990

Eder, A. B., & Hommel, B. (2013). Anticipatory control of approach and avoidance: An ideomotor approach. *Emotion Review*, *5*(3), 275–279. https://doi.org/10.1177/1754073913477505

Eder, A. B., & Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology: General*, *137*(2), 262–281. https://doi.org/10.1037/0096-3445.137.2.262

Ellwart, T., Rinck, M., & Becker, E. S. (2006). From fear to love: individual differences in implicit spider associations. *Emotion (Washington, D.C.)*, *6*(1), 18–27. https://doi.org/10.1037/1528-3542.6.1.18

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238.

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

Hofmann, W., Friese, M., & Roefs, A. (2009). Three ways to resist temptation: The independent contributions of executive attention, inhibitory control, and affect regulation to the impulse control of eating behavior. *Journal of Experimental Social Psychology*, *45*(2), 431–435. https://doi.org/10.1016/j.jesp.2008.09.013

Hong, S.-M., & Page, S. (1989). A Psychological Reactance Scale: Development, Factor Structure and Reliability. *Psychological Reports*, *64*(3), 1323–1326. https://doi.org/10.2466/pr0.1989.64.3c.1323

Huijding, J., Field, A. P., De Houwer, J., Vandenbosch, K., Rinck, M., & van Oeveren, M. (2009). A behavioral route to dysfunctional representations: The effects of training approach or avoidance tendencies towards novel animals in children. *Behaviour Research and Therapy*, *47*(6), 471–477. https://doi.org/10.1016/j.brat.2009.02.011

Jones, C. R., Vilensky, M. R., Vasey, M. W., & Fazio, R. H. (2013). Approach behavior can mitigate predominately univalent negative attitudes: Evidence regarding insects and spiders. *Emotion*, *13*(5), 989–996. https://doi.org/10.1037/a0033164

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, *77*(5), 361–372. https://doi.org/10.1037/h0032590

Kassem, N. O., Lee, J. W., Modeste, N. N., & Johnston, P. K. (2003). Understanding soft drink consumption among female adolescents using the Theory of Planned Behavior. *Health Education Research*, *18*(3), 278–291. https://doi.org/10.1093/her/cyf017

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder:Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, *92*(6), 957–971. https://doi.org/10.1037/0022-3514.92.6.957

Krishna, A., & Eder, A. B. (in press). No effects of approach-avoidance training on consumption of soft drinks. *Appetite.*

Krypotos, A.-M., Arnaudova, I., Effting, M., Kindt, M., & Beckers, T. (2015). Effects of Approach-Avoidance Training on the Extinction and Return of Fear Responses. *PLOS ONE*, *10*(7), e0131581. https://doi.org/10.1371/journal.pone.0131581

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition and Emotion*, *23*(3), 453–480. https://doi.org/10.1080/02699930801994054

Lindgren, K. P., Wiers, R. W., Teachman, B. A., Gasser, M. L., Westgate, E. C., Cousijn, J., … Neighbors, C. (2015). Attempted Training of Alcohol Approach and Drinking

Identity Associations in US Undergraduate Drinkers: Null Results from Two Studies. *PLOS ONE*, *10*(8), e0134642. https://doi.org/10.1371/journal.pone.0134642

Mertens, G., Van Dessel, P., & De Houwer, J. (2018). The contextual malleability of approach-avoidance training effects: approaching or avoiding fear conditioned stimuli modulates effects of approach-avoidance training. *Cognition and Emotion*, *32*(2), 341–349. https://doi.org/10.1080/02699931.2017.1308315

Paulus, A., & Wentura, D. (2016). It depends: Approach and avoidance reactions to emotional expressions are influenced by the contrast emotions presented in the task. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 197–212. https://doi.org/10.1037/xhp0000130

Perugini, M., Richetin, J., & Zogmaister, C. (2012). The formation of implicit and explicit attitudes for neutral and valenced stimuli using the self. *Learning and Motivation*, *43*(3), 135–143. https://doi.org/10.1016/j.lmot.2012.02.002

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, *61*(1), 37–54. https://doi.org/10.1016/j.jml.2009.03.004

Stritzke, W. G. K., McEvoy, P. M., Wheat, L. R., Dyer, K. R., & French, D. J. (2007). The Yin and Yang of Indulgence and Restraint: The Ambivalence Model of Craving. In P. W. O'Neal (Ed.), *Motivation of health behavior* (pp. 31–47). New York: Nova Science Publishers.

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let's not be indifferent about (attitudinal) ambivalence. In *Attitude strength: Antecedents and consequences.* (pp. 361–386). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Van Dessel, P., De Houwer, J., & Gast, A. (2016). Approach–Avoidance Training Effects Are Moderated by Awareness of Stimulus–Action Contingencies. *Personality and Social Psychology Bulletin*, *42*(1), 81–93. https://doi.org/10.1177/0146167215615335

Van Dessel, P., De Houwer, J., Roets, A., & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. *Journal of Personality and Social Psychology*, *110*(1), e1–e15. https://doi.org/10.1037/pspa0000039

Van Dessel, P., Eder, A., & Hughes, S. J. (2018). Mechanisms underlying effects of approach-avoidance training on stimulus evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000514

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*, *Complete*(69), 23–32. https://doi.org/10.1016/j.jesp.2016.10.004

Van Dessel, P., Hughes, S., & De Houwer, J. (in press). How Do Actions Influence Attitudes? An Inferential Account of the Impact of Action Performance on Stimulus Evaluation. *Personality and Social Psychology Review*. Retrieved from https://biblio.ugent.be/publication/8570657/file/8570659

van Uijen, S. L., van den Hout, M., & Engelhard, I. M. (2015). Active Approach Does not Add to the Effects of in Vivo Exposure. *Journal of Experimental Psychopathology*, *6*(1), 112–125. https://doi.org/10.5127/jep.042014

Vandenbosch, K., & De Houwer, J. (2011). Failures to induce implicit evaluations by means of approach–avoid training. *Cognition and Emotion*, *25*(7), 1311–1330. https://doi.org/10.1080/02699931.2011.596819

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining Automatic Action Tendencies Changes Alcoholic Patients' Approach Bias for Alcohol and Improves Treatment Outcome. *Psychological Science*, *22*(4), 490–497. https://doi.org/10.1177/0956797611400615

Wiers, R. W., Houben, K., Fadardi, J. S., van Beek, P., Rhemtulla, M., & Cox, W. M. (2015). Alcohol Cognitive Bias Modification training for problem drinkers over the web. *Addictive Behaviors*, *40*, 21–26. https://doi.org/10.1016/j.addbeh.2014.08.010

Wiers, R. W., Rinck, M., Kordts, R., Houben, K., & Strack, F. (2010). Retraining automatic action-tendencies to approach alcohol in hazardous drinkers. *Addiction*, *105*(2), 279–287. https://doi.org/10.1111/j.1360-0443.2009.02775.x

Wilkowski, B. M., & Meier, B. P. (2010). Bring it on: Angry facial expressions potentiate approach-motivated motor behavior. *Journal of Personality and Social Psychology*, *98*(2), 201–210.

Woud, M. L., Becker, E. S., Lange, W.-G., & Rinck, M. (2013). Effects of Approach-Avoidance Training on Implicit and Explicit Evaluations of Neutral, Angry, and Smiling Face Stimuli. *Psychological Reports*, *113*(1), 199–216. https://doi.org/10.2466/21.07.PR0.113x10z1

Zogmaister, C., Perugini, M., & Richetin, J. (2016). Motivation modulates the effect of approach on implicit preferences. *Cognition and Emotion*, *30*(5), 890–911. https://doi.org/10.1080/02699931.2015.1032892

Tables

Table 1

Descriptives of pre-training evaluative measures in Experiment 1

	Mezzo Mix	Schwip Schwap
	M (SD)	*M (SD)*
Positive components	2.86 (.87)	2.90 (.87)
Negative components	2.81 (.92)	2.68 (.97)
Average certainty score	6.45 (1.73)	6.51 (1.73)
Implicit liking (RT)	21.4 (46.8)	9.1 (42.9)
Implicit liking (Error)	-0.4% (6.1)	-0.3% (7.1)
Ambivalence score	1.60 (1.06)	1.64 (1.19)

Note. Positive and negative components were scored on scales from 1 to 4. Certainty ratings were scored on a scale from 1 to 9. Ambivalence scores could range from -0.5 to 4. No ratings differed between drinks (all *ps* > .20) except implicit liking RT (p = .062).

Table 2

Descriptives of pre-training evaluative measures, Experiment 2

	Mezzo Mix	Schwip Schwap
	M (SD)	*M (SD)*
Positive components	1.63 (1.20)	1.68 (1.09)
Negative components	1.15 (1.13)	1.25 (1.11)
Certainty score	6.73 (1.76)	6.75 (1.78)
Implicit liking (RT)	10.5 (40.2)	13.7 (50.0)
Implicit liking (Error)	-0.5% (8.4)	-1.0% (6.6)
Ambivalence	.03 (1.15)	.09 (1.13)

Note. Positive and negative components were scored on the evaluative grid from 0 to 4. Certainty ratings were scored on a scale from 1 to 9. Ambivalence scores could range from -2 to 4.

Figures

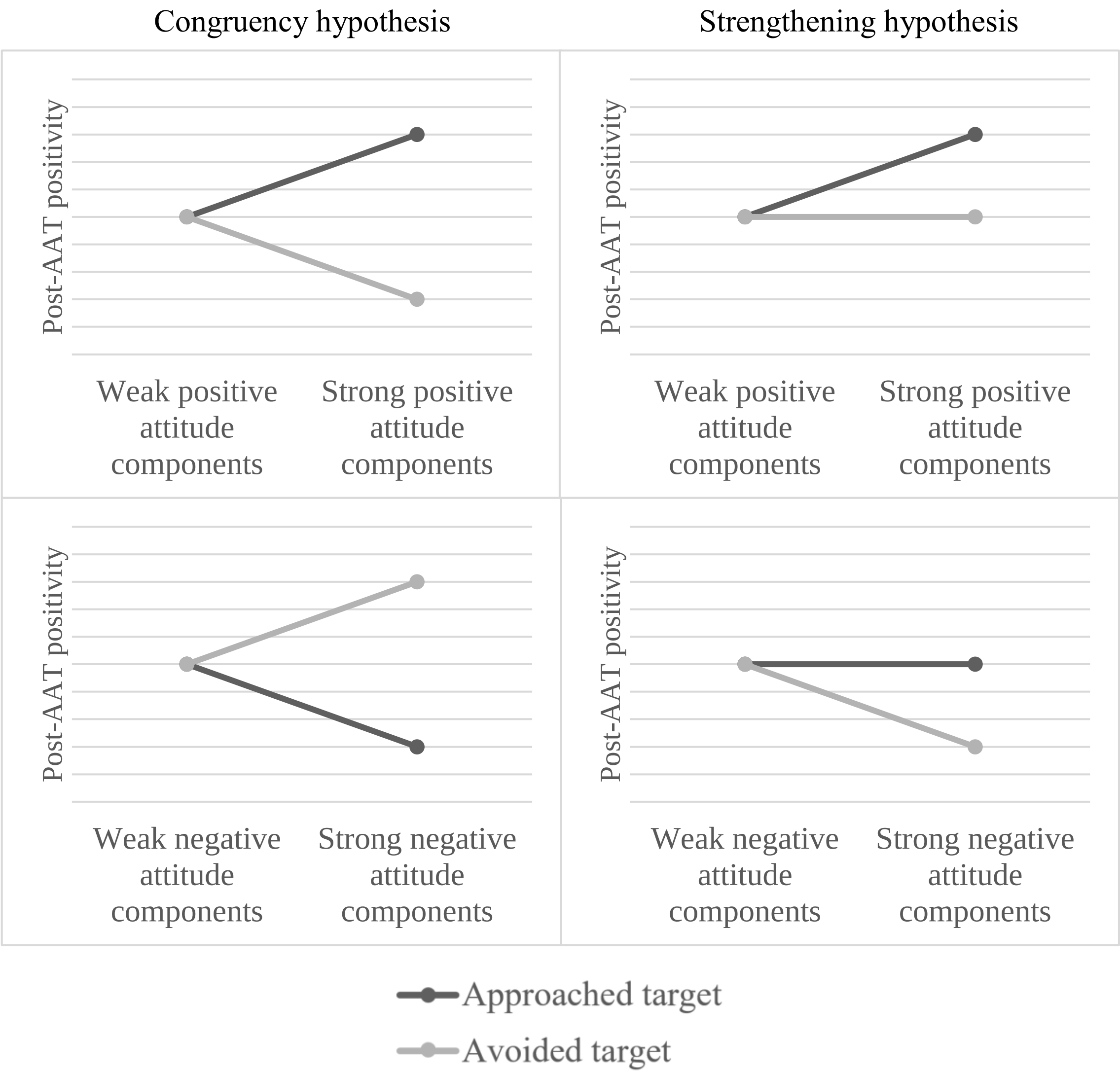

Figure 1

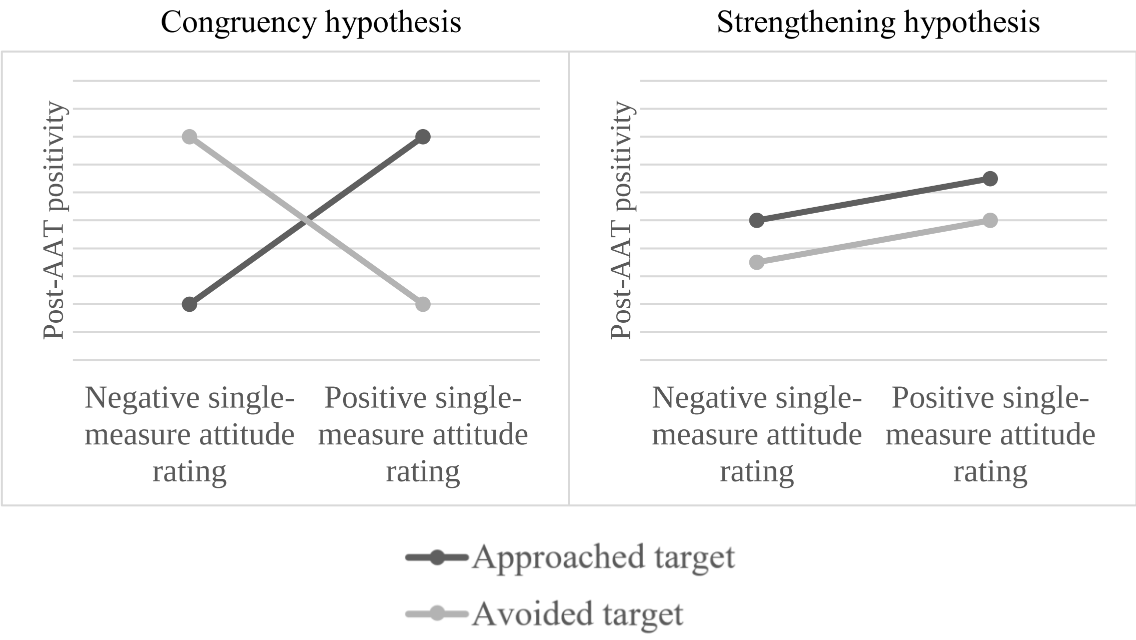

Figure 2

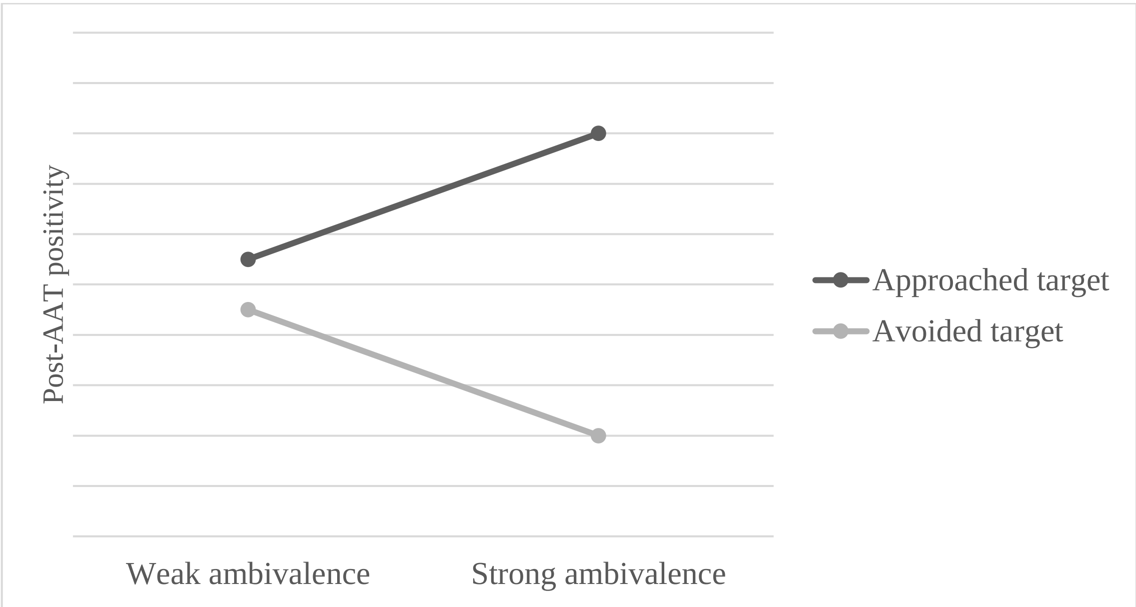

Figure 3

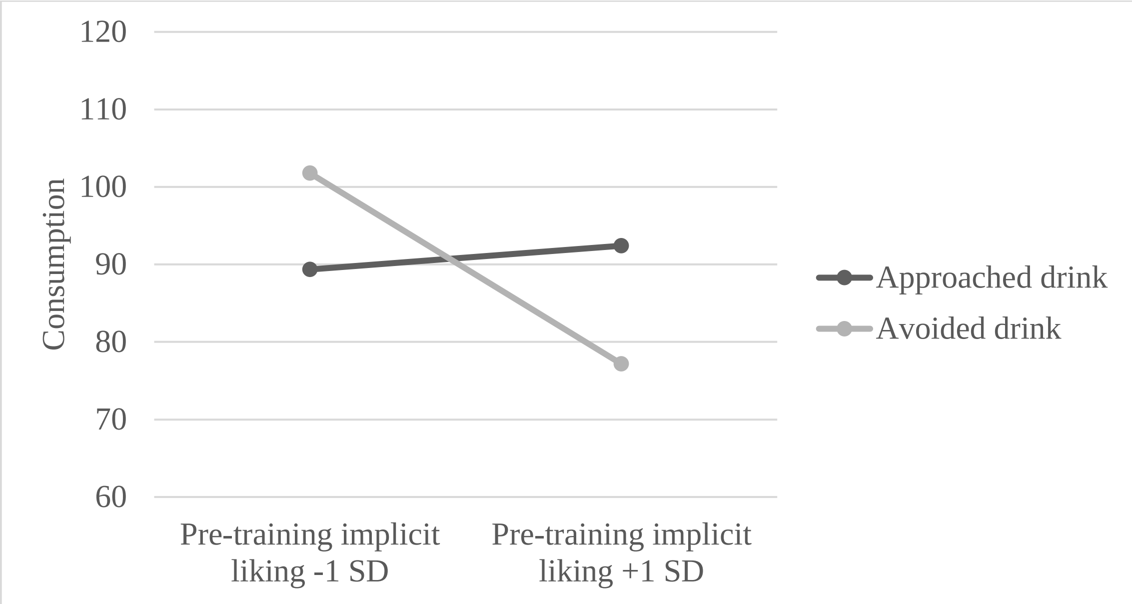

Figure 4

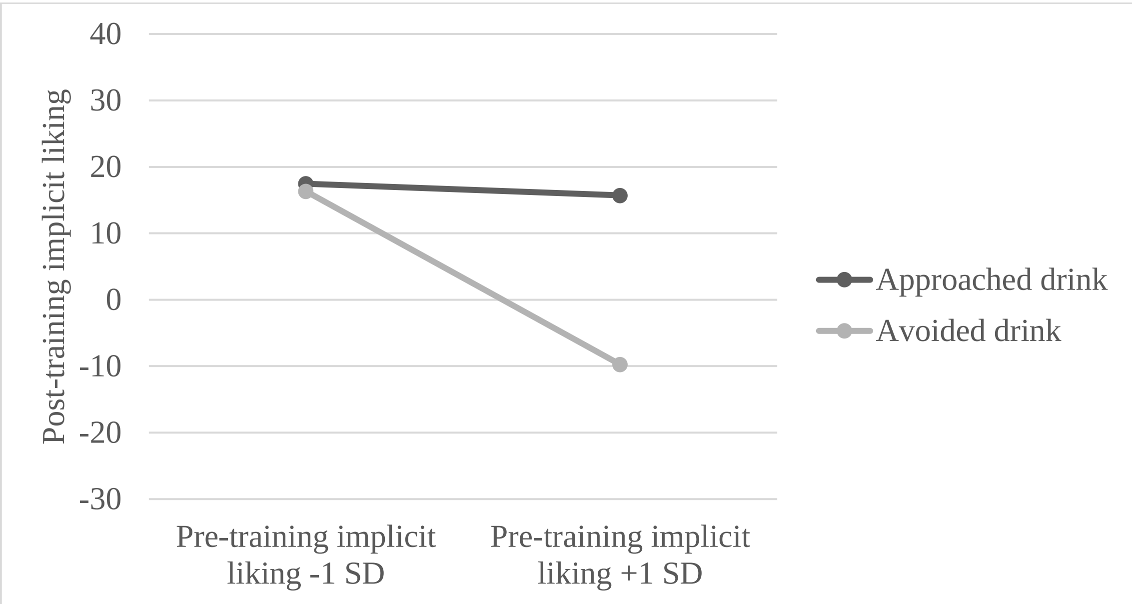

Figure 5

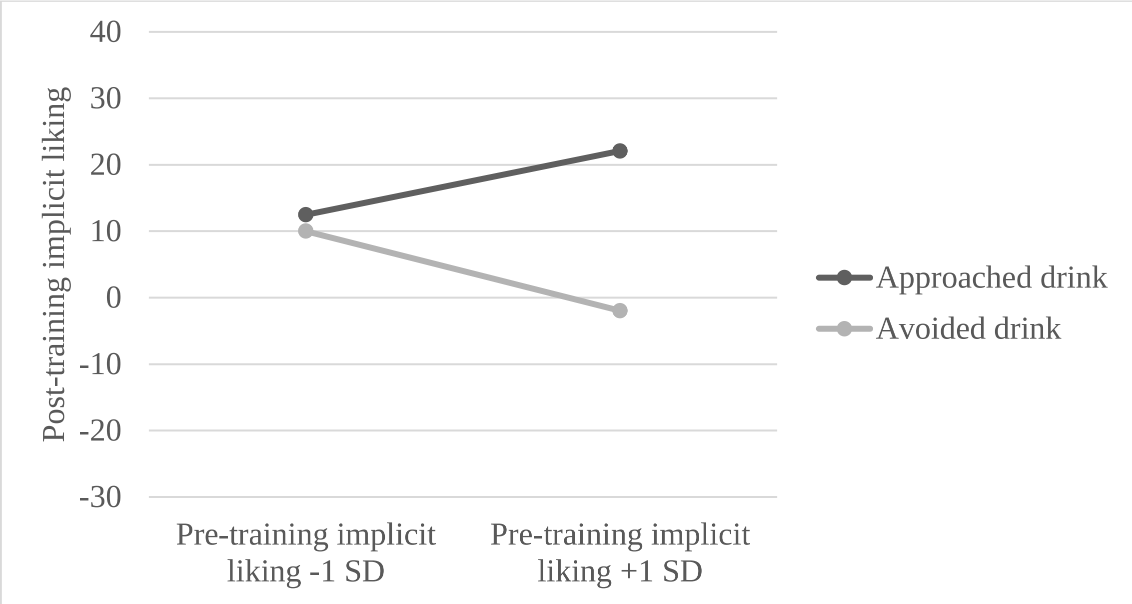

Figure 6

Figure Captions

Figure 1. Hypothesized patterns for the congruency and strengthening hypotheses for pre-training explicit evaluative components. The Y-axis refers to post-training consumption (higher values imply more consumption) and explicit ratings (higher values imply more positive ratings).

Figure 2. Hypothesized patterns for the congruency and strengthening hypotheses for pre-training implicit positivity. The Y-axis refers to post-training consumption (higher values imply more consumption) and implicit preferences (higher values imply more positive values).

Figure 2. Hypothesized pattern for the ambivalence hypothesis. The Y-axis refers to post-training consumption (higher values imply more consumption), explicit ratings (higher values imply more positive ratings) and implicit preferences (higher values imply more positive values).

Figure 4. Model estimates for drink consumption as a function of AAT and prior implicit liking in Experiment 1. The unit of measurement is grams.

Figure 5. Model estimates for post-training implicit positivity as a function of AAT and prior implicit liking in Experiment 1.

Figure 6. Model estimates for post-training implicit positivity as a function of AAT and prior implicit liking in Experiment 1 and 2 (mega-analysis).

Supplementary Materials

Targets for affective priming task

Target words for the affective priming task were selected from the BAWL-R database (Võ et al., 2009) according to their valence and matched for length and frequency (both |*ts*| < 1). Negative targets with bracketed English translations were: MIES (lousy), TRIST (forlorn), BRUTAL (sadistic), WEHRLOS (defenceless), BANKROTT (bankrupt), KRAFTLOS (feeble), MILITANT (militant), TROSTLOS (cheerless), ENTSETZT (horrified), and VERBOTEN (forbidden); positive targets were: TOLL (swell), SUPER (superb), SONNIG (sunny), PFIFFIG (gutsy), TAKTVOLL (tactful), TOLERANT (tolerant), LUKRATIV (lucrative), GRANDIOS (great), SINNLICH (sensual), and REIZVOLL (attractive).

Outlier elimination rules

In both experiments, participants were excluded from all analyses if their error rate in the AAT task exceeded the third quartile of the experiment's sample by three interquartiles (far outlier criterion; Tukey, 1977). The criterion for Experiment 1 was 9.5% errors (two eliminated), the criterion for Experiment 2 was 12.5% errors (one eliminated).

Furthermore, in both experiments, participants were eliminated from any analyses including indices derived from the affective priming tasks if their error rate in either the pre- or post-training affective priming exceeded a sample-level far outlier criterion. In Experiment 1, the criterion was 19.6% errors pre-training (one excluded) and 23.9% errors post-training (none excluded). In Experiment 2, the criterion was 27.2% errors pre-training (one excluded) and 22.8% errors post-training (none excluded).

In both experiments, for both the affective priming tasks and the behavioral assessment tasks, trials with an incorrect response were eliminated from analysis. In Experiment 1, this was 4.0% of trials in the behavioral assessment task, 4.7% of trials in the pre-training affective priming and 5.3% of trials in the post-training affective priming. In Experiment 2, it was 5.2% of trials in the behavioral assessment task, 7.4% of trials in the pre-training affective priming and 6.3% of trials in the post-training affective priming.

Finally, in both experiments, for both the affective priming tasks and the behavioral assessment tasks, trials with a RT below 100ms or above an individual simple outlier criterion (exceeding their third quartile by at least 1.5 interquartiles) were eliminated. In Experiment 1, this was 5.4% of remaining trials in the behavioral assessment task, 5.1% of remaining trials in the pre-training affective priming and 4.6% of remaining trials in the post-training affective priming. In Experiment 2, it was 4.7% of remaining trials in the behavioral assessment task, 4.8% of remaining trials in the pre-training affective priming and 4.8% of remaining trials in the post-training affective priming.

Moderator analyses

The models calculated for the mega-analysis were subjected to moderator analyses by including thirst ratings, mood rating, trait and state reactance ratings as well as all their respective interaction terms with the other predictors in separate models for each moderator. Due to the explorative nature of these analyses and the number of tests calculated, only models containing novel effects with $p < .01$ are reported in order to reduce α errors. Full tables are omitted for brevity; analysis scripts may be found under http://osf.io/t7kmf.

Thirst. Thirst has no significant moderation effect in models predicting consumption or implicit liking, but in those (explicit attitude components and explicit ambivalence) for explicit ratings, the interaction of thirst and AAT achieves significance, $F(1,122.5) = 7.25$, $p = .008$ (explicit attitude components model). Thirst tends to decrease liking for the avoided drink, $B = -.16$, $p = .197$, but increases it for the approached drink, $B = .27$, $p = .034$ (see Figure 1).

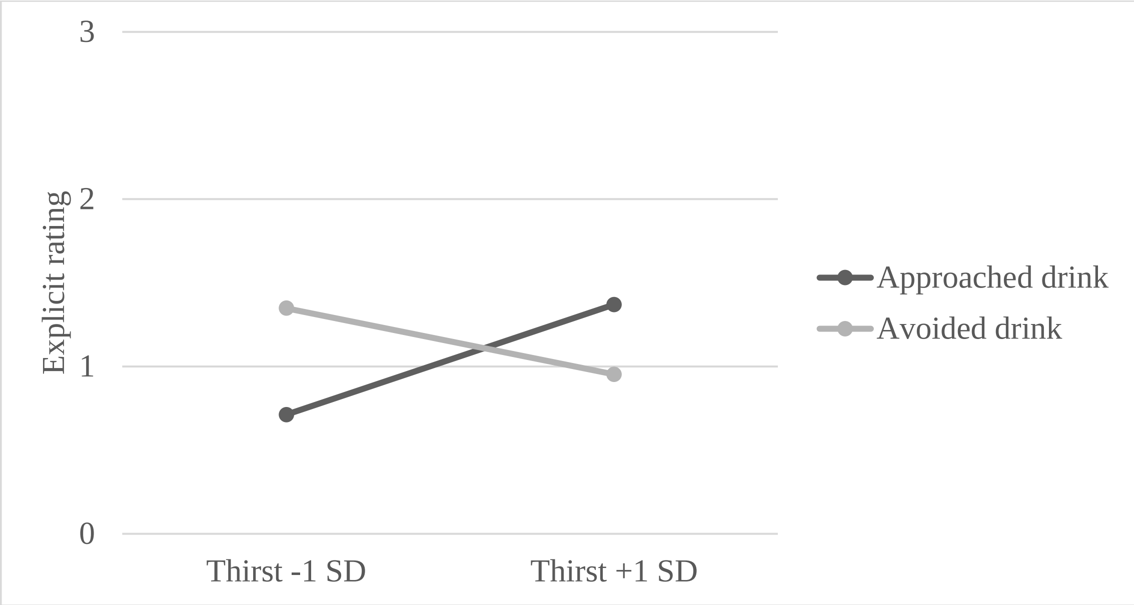

Figure 1: Model estimates for explicit rating as a function of AAT and thirst, mega-analysis of Experiments 1 and 2.

Our results agree with those of Zogmaister and colleagues (2016) in that a relevant deprivation state increases AAT effects, but differ from them in that this occurs only on explicit measures in our studies, but only on implicit measures in theirs. Furthermore, they find the opposite interaction for their explicit measure (a reduction of AAT effects with increasing thirst) in their only experiment concerning soft drinks. However, there are two important differences between our studies that may explain this discrepancy: First, Zogmaister et al.'s soft drink experiment did not contain an AAT, but rather tested approach training against behavioral inhibition. Although inhibition has been linked to avoidance, they address separate neural systems (Amodio, Master, Yee, & Taylor, 2007) and may diverge in training tasks. In Zogmaister et al.'s other experiments, they used avoidance rather than inhibition training and found no opposed interaction. Second, in our experiments, participants had the opportunity to satiate their thirst before the explicit rating. Therefore, participants may have been thirsty during the training phase, but were not during the explicit attitude measurement. As deprivation states have been shown to modulate affective responses to satiation-relevant stimuli (Hoefling et al., 2009), there is reason to assume that participants' judgments might have been affected by their thirst in Zogmaister and colleagues' work. For example, participants might have noted a stronger spontaneous emotional response to the approach-trained drinks, but attributed this to their thirst or discounted its relevance in their propositional response (Gawronski & Bodenhausen, 2006), thereby producing reversed or no effects on the explicit measures. Alternatively, their frustration at judging a drink, but not being able to actually drink it while thirsty may have mitigated any explicit training effect. For these important procedural reasons, our study should not be considered a failed replication of Zogmaister and colleagues. Instead, these differences may provide insight into the processes involved in motivational moderation of AAT: Zogmaister and colleagues' results reflect greater AAT effects on spontaneous evaluations when in a state of deprivation. In contrast, in our experiments, we show that participants' motivational state during the training phase affects explicit judgments even after that motivational state is mitigated, implying an encoding-level effect of motivational states.

Reactance. Although state reactance showed no moderation effects or other interactions in any model by our criteria, trait reactance produced an unexpected three-way interaction effect with ambivalence and AAT in the model predicting explicit ratings, $F(1,232.2) = 6.64$, $p = .002$ (see Figure 2).

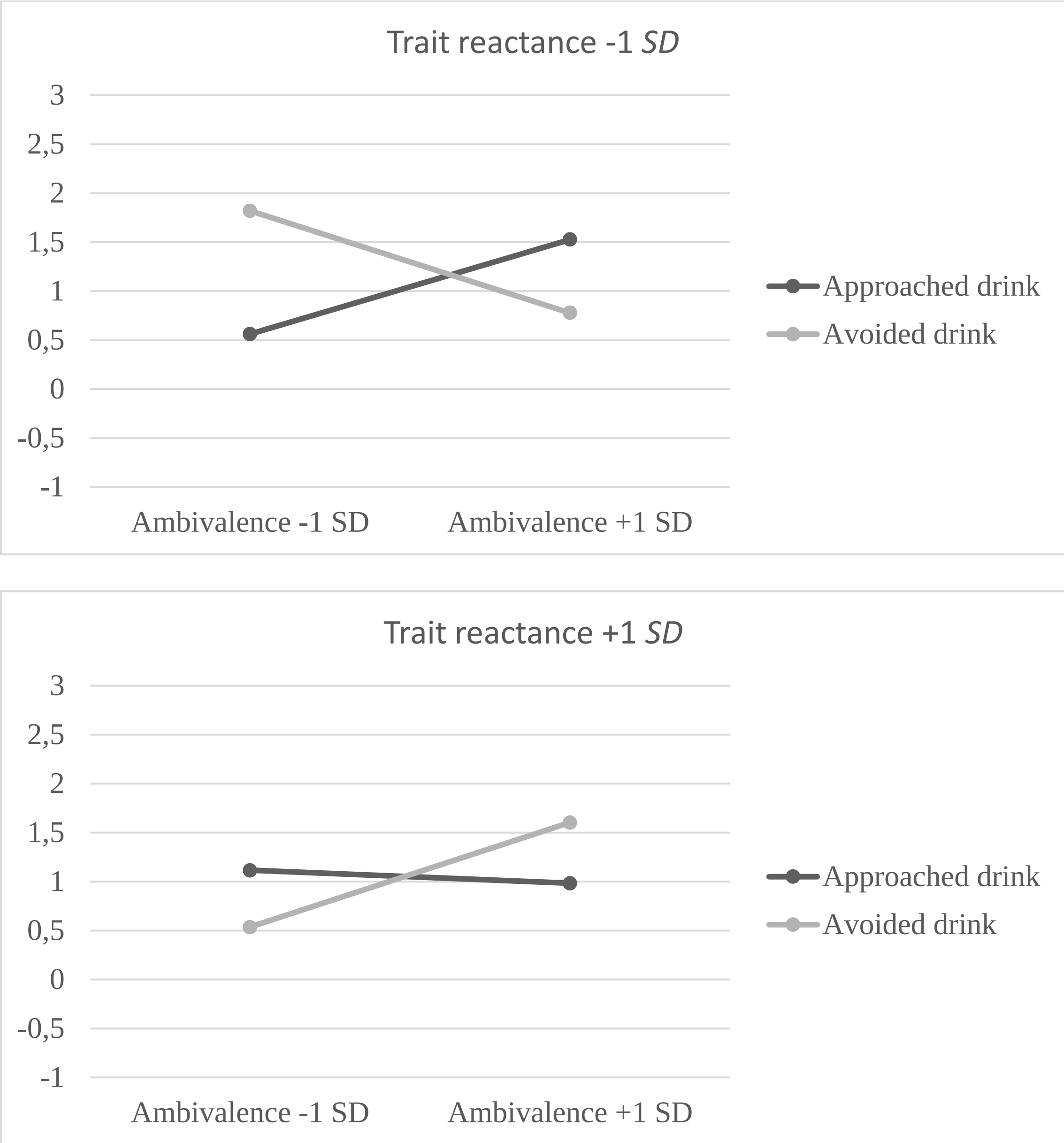

Figure 2: Model estimates for explicit rating as a function of AAT, ambivalence and trait reactance, mega-analysis of Experiments 1 and 2.

Reactance effects occur when participants believe they have understood the expected effects of experimental manipulations and act against these expectations in order to assert their freedom (Brehm, 1966). Therefore, no reactance effect should be interpreted as evidence for the psychological mechanisms under scrutiny. Instead, reactance effects must be viewed from this perspective of motivated resistance to influence. In our study, greater pre-training ambivalence led to greater AAT effects on explicit attitudes in participants with low trait reactance, but the opposite effect occurred in participants with high trait reactance. This may be an indication that ambivalence plays a role in AAT effects, but that its effect is partially or wholly mediated by conscious processes. For example, approach training might affect ambivalent attitudes by selectively increasing the accessibility of positive attitude components with regard to the target via a general approach orientation (Strack & Deutsch, 2004), with univalent attitudes being less affected due to ceiling effects (Woud, Becker, & Rinck, 2011). For low-reactance individuals, this increase in accessibility leads to more positive contents being integrated into the explicit judgment, but for high-reactance individuals, the activated contents might be dismissed as irrelevant for the judgment due to a perception that they are caused by experimental manipulation, leading to reversed AAT effects (Gawronski & Bodenhausen, 2006). Future research might therefore integrate ambivalence as a predictor of AAT effects, especially in unobtrusive AAT paradigms where reactance is less likely. On a more general level, this pattern of results underlines the importance of considering reactance effects for explicit judgments in AAT paradigms and offers a tentative suggestion as to why AAT studies may sometimes produce effects only on implicit measures (e.g. Woud, Maas, Becker, & Rinck, 2013), one which may be investigated in future studies.

Tables for mixed-model analyses

The following tables describe the results of all mixed-model analyses completely.

Experiment 1.

Table 3

Consumption, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	137.4	10.07	.002
AAT	68.1	.02	.888
Positive attitude	68.2	3.82	.055
Negative attitude	68.2	.01	.938
Positive attitude (subject mean)	137.8	3.61	.059
Negative attitude (subject mean)	135.2	.34	.563
AAT*Positive attitude	74.3	.12	.730
AAT*Negative attitude	76.9	.00	.954

Note. Numerator df was 1 for all tests.

Table 4

Consumption, implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	77.6	132.48	<.001
AAT	69.2	.05	.823
Implicit positivity	69.2	1.13	.291
Implicit positivity (subject mean)	112.6	.75	.387
AAT*Implicit positivity	87.8	3.35	.070

Note. Numerator df was 1 for all tests.

Table 5

Consumption, implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	70.2	180.80	<.001
AAT	69.2	.02	.903
Implicit positivity	69.3	1.78	.186
Implicit positivity (subject mean)	111.2	2.04	.156
AAT*Implicit positivity	89.7	.50	.480

Note. Numerator df was 1 for all tests.

Table 6

Consumption, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	71.4	180.27	<.001
AAT	70.4	.01	.919
Ambivalence	132.1	.01	.910
AAT*Ambivalence	82.8	.13	.716

Note. Numerator df was 1 for all tests.

Table 7

Post-training implicit positivity (RT), implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	83.3	.21	.651
AAT	68.1	6.26	.015
Implicit positivity	68.1	2.66	.107
Implicit positivity (subject mean)	133.7	11.68	.001
AAT*Implicit positivity	102.9	4.07	.046

Note. Numerator df was 1 for all tests.

Table 8

Post-training implicit positivity (error), implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	70.6	2.82	.098
AAT	69.4	.05	.818
Implicit positivity	69.5	.31	.581
Implicit positivity (subject mean)	138.5	1.36	.246
AAT*Implicit positivity	126.3	.12	.732

Note. Numerator df was 1 for all tests.

Table 9

Post-training implicit positivity (RT), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	71.4	6.51	.013
AAT	70.8	4.78	.032
Ambivalence	141.8	.00	.960
AAT*Ambivalence	93.9	1.09	.299

Note. Numerator df was 1 for all tests.

Table 10

Post-training implicit positivity (error), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	71.3	1.45	.233
AAT	70.9	.06	.803
Ambivalence	137.2	.02	.898
AAT*Ambivalence	101.3	1.40	.240

Note. Numerator df was 1 for all tests.

Table 11

Post-training explicit rating, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	138	1.33	.250
AAT	138	.43	.515
Positive attitude	138	4.55	.035
Negative attitude	138	.08	.772
Positive attitude (subject mean)	138	.00	.996
Negative attitude (subject mean)	138	1.02	.314
AAT*Positive attitude	138	.36	.552
AAT*Negative attitude	138	.58	.449

Note. Numerator df was 1 for all tests.

Table 12

Post-training explicit rating, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	71.4	49.41	<.001
AAT	71.2	.47	.495
Ambivalence	130.5	.30	.585
AAT*Ambivalence	108.5	.11	.746

Note. Numerator df was 1 for all tests.

Experiment 2.

Table 13

Consumption, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	105.5	12.23	.001
AAT	53.1	.06	.811
Positive attitude	53.2	.65	.422
Negative attitude	53.1	.05	.821
Positive attitude (subject mean)	107.9	.15	.697
Negative attitude (subject mean)	107.7	.69	.408
AAT*Positive attitude	56.0	3.59	.063
AAT*Negative attitude	56.1	1.52	.223

Note. Numerator df was 1 for all tests.

Table 14

Consumption, implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	58.1	96.19	<.001
AAT	54.1	.12	.727
Implicit positivity	54.2	.93	.340
Implicit positivity (subject mean)	77.1	.15	.697
AAT*Implicit positivity	65.3	.17	.681

Note. Numerator df was 1 for all tests.

Table 15

Consumption, implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	55.6	105.30	.000
AAT	54.1	.00	.993
Implicit positivity	54.1	.10	.757
Implicit positivity (subject mean)	76.8	.22	.643
AAT*Implicit positivity	65.4	2.23	.140

Note. Numerator df was 1 for all tests.

Table 16

Consumption, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	57.2	114.11	<.001
AAT	57.7	.01	.922
Ambivalence	101.8	.14	.708
AAT*Ambivalence	60.5	.18	.676

Note. Numerator df was 1 for all tests.

Table 17

Post-training implicit positivity (RT), implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	70.1	4.15	.046
AAT	55.6	7.32	.009
Implicit positivity	55.7	2.53	.117
Implicit positivity (subject mean)	110.9	3.61	.060
AAT*Implicit positivity	99.8	2.49	.118

Note. Numerator df was 1 for all tests.

Table 18

Post-training implicit positivity (error), implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	58.1	2.83	.098
AAT	54.6	1.71	.196
Implicit positivity	54.8	.33	.570
Implicit positivity (subject mean)	109.5	2.76	.100
AAT*Implicit positivity	104.7	0.59	.446

Note. Numerator df was 1 for all tests.

Table 19

Post-training implicit positivity (RT), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	58.3	7.00	.010
AAT	58.8	3.82	.055
Ambivalence	112.5	.94	.333
AAT*Ambivalence	82.7	.00	.957

Note. Numerator df was 1 for all tests.

Table 20

Post-training implicit positivity (error), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	59.6	3.92	.052
AAT	59.7	.66	.419
Ambivalence	104.6	1.36	.246
AAT*Ambivalence	93.8	.94	.334

Note. Numerator df was 1 for all tests.

Table 21

Post-training explicit rating, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	96.8	1.15	.287
AAT	53.7	.36	.551
Positive attitude	54.6	1.50	.226
Negative attitude	53.7	7.67	.008
Positive attitude (subject mean)	94.2	2.09	.151
Negative attitude (subject mean)	83.6	.94	.334
AAT*Positive attitude	66.1	.07	.786
AAT*Negative attitude	63.5	.01	.926

Note. Numerator df was 1 for all tests.

Table 22

Post-training explicit rating, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	58.2	16.71	<.001
AAT	58.6	.03	.857
Ambivalence	103.5	.02	.890
AAT*Ambivalence	67.7	.15	.697

Note. Numerator df was 1 for all tests.

Mega-analysis.

Table 23

Consumption, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	127.9	275.60	<.001
AAT	126.0	.03	.863
Positive attitude	126.1	4.16	.043
Negative attitude	126.0	.04	.836
Positive attitude (subject mean)	253.3	1.14	.287
Negative attitude (subject mean)	246.7	1.08	.300
AAT*Positive attitude	134.4	2.16	.144
AAT*Negative attitude	136.6	.75	.387

Note. Numerator df was 1 for all tests.

Table 24

Consumption, implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	135.9	220.89	<.001
AAT	126.1	.03	.853
Implicit positivity	126.1	.01	.924
Implicit positivity (subject mean)	187.3	.02	.875
AAT*Implicit positivity	155.2	.82	.365

Note. Numerator df was 1 for all tests.

Table 25

Consumption, implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	127.7	265.34	<.001
AAT	126.2	.04	.840
Implicit positivity	126.2	.41	.521
Implicit positivity (subject mean)	185.6	.15	.695
AAT*Implicit positivity	156.0	2.51	.115

Note. Numerator df was 1 for all tests.

Table 26

Consumption, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	129.9	273.06	<.001
AAT	128.7	.01	.935
Ambivalence	231.7	.08	.772
AAT*Ambivalence	144.3	.01	.926

Note. Numerator df was 1 for all tests.

Table 27

Post-training implicit positivity (RT), implicit positivity RT model

Parameter	Denominator *df*	*F*	*p*
Intercept	155.0	2.62	.107
AAT	126.4	10.95	.001
Implicit positivity	126.5	.04	.852
Implicit positivity (subject mean)	251.8	14.89	<.001
AAT*Implicit positivity	208.6	5.89	.016

Note. Numerator df was 1 for all tests.

Table 28

Post-training implicit positivity (error), implicit positivity error model

Parameter	Denominator *df*	*F*	*p*
Intercept	130.3	6.09	.015
AAT	126.6	.54	.464
Implicit positivity	126.8	.85	.357
Implicit positivity (subject mean)	253.2	3.73	.054
AAT*Implicit positivity	237.3	.14	.706

Note. Numerator df was 1 for all tests.

Table 29

Post-training implicit positivity (RT), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	129.4	13.59	<.001
AAT	128.8	9.70	.002
Ambivalence	258.6	.37	.546
AAT*Ambivalence	178.5	.60	.439

Note. Numerator df was 1 for all tests.

Table 30

Post-training implicit positivity (error), ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	129.9	5.28	.023
AAT	129.6	.32	.575
Ambivalence	245.6	.73	.394
AAT*Ambivalence	197.3	2.35	.127

Note. Numerator df was 1 for all tests.

Table 31

Post-training explicit rating, explicit attitude components model

Parameter	Denominator *df*	*F*	*p*
Intercept	128.7	104.75	<.001
AAT	127.3	.49	.486
Positive attitude	128.1	5.75	.018
Negative attitude	127.3	1.11	.295
Positive attitude (subject mean)	197.4	.55	.457
Negative attitude (subject mean)	199.4	3.40	.067
AAT*Positive attitude	167.8	.18	.674
AAT*Negative attitude	170.7	1.00	.318

Note. Numerator df was 1 for all tests.

Table 32

Post-training explicit rating, ambivalence model

Parameter	Denominator *df*	*F*	*p*
Intercept	129.9	61.29	<.001
AAT	129.2	.47	.496
Ambivalence	259.4	.40	.527
AAT*Ambivalence	176.5	.17	.679

Note. Numerator df was 1 for all tests.

References for Supplementary Materials

Amodio, D. M., Master, S. L., Yee, C. M., & Taylor, S. E. (2007). Neurocognitive components of the behavioral inhibition and activation systems: Implications for theories of self-regulation. *Psychophysiology*, *45*(1), 11–19. https://doi.org/10.1111/j.1469-8986.2007.00609.x

Brehm, J. W. (1966). *A theory of psychological reactance.* Oxford, England: Academic Press.

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

Hoefling, A., Likowski, K. U., Deutsch, R., Häfner, M., Seibt, B., Mühlberger, A., … Strack, F. (2009). When hunger finds no fault with moldy corn: food deprivation reduces food-related disgust. *Emotion (Washington, D.C.)*, *9*(1), 50–58. https://doi.org/10.1037/a0014449

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Tukey, J. (1977). *Exploratory Data Analysis*. Pearson. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0201076160

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Woud, M. L., Becker, E. S., & Rinck, M. (2011). Induction of implicit evaluation biases by approach–avoidance training: A commentary on Vandenbosch and De Houwer (this issue). *Cognition & Emotion*, *25*(7), 1331–1338. https://doi.org/10.1080/02699931.2011.628300

Woud, M. L., Maas, J., Becker, E. S., & Rinck, M. (2013). Make the manikin move: Symbolic approach–avoidance responses affect implicit and explicit face evaluations. *Journal of Cognitive Psychology*, *25*(6), 738–744. https://doi.org/10.1080/20445911.2013.817413

Zogmaister, C., Perugini, M., & Richetin, J. (2016). Motivation modulates the effect of approach on implicit preferences. *Cognition and Emotion*, *30*(5), 890–911. https://doi.org/10.1080/02699931.2015.1032892