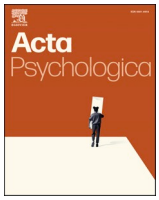
Acta Psychologica 223 (2022) 103485

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Effects of self-instructed stimulus-affect plans on indirectly measured and self-reported evaluative responses☆

Torsten Martiny-Huenger [a,*], Jenny Roth [b,c,**]

[a] UiT The Arctic University of Norway, Norway
[b] University of Limerick, Limerick, Ireland
[c] University of California Davis, CA, USA

ARTICLE INFO

Keywords:
Stimulus-response learning
Implementation intentions
Attitude change
Implicit association test

ABSTRACT

Repeatedly experiencing a specific stimulus-affect contingency influences subsequent evaluative responses towards the respective stimulus (e.g., evaluative conditioning). In the present research, we provide further evidence that verbally processed stimulus-affect contingencies in the form of if-then plans have comparable evaluative consequences. We present three studies ($N = 323$) in which participants verbally linked cupcakes to either a positive ("delicious") or a negative ("disgusting") affective response while being instructed with the same health-related goal. We tested the evaluative consequences of processing these verbal stimulus-affect plans in a valence-based response-compatibility paradigm (Implicit Association Test, IAT) and self-reported liking ratings. We failed to observe the predicted effect in the first study and updated the methodology for the following two studies. With the updated procedure (two studies, $N = 239$), we found the hypothesized effect that processing a verbal stimulus-affect plan influences subsequent responses in the IAT and self-reported ratings in an evaluatively congruent direction. We discuss these results in relation to similar effects following directly experienced stimulus-affect contingencies and instructed evaluative conditioning. Furthermore, our present research highlights the potential to use verbal self-instruction in a stimulus-affect format to self-regulate one's evaluative responses towards specific stimuli (e.g., unhealthy snacks).

1. Effects of self-instructed stimulus-affect plans on indirectly measured and self-reported evaluative responses

Direct experiences influence subsequent evaluative and behavioral responses (i.e., learning; e.g., Pavlov, 1927). For example, repeatedly performing a behavior in a particular context leads to a more efficient initiation of the behavior in the same context (e.g., stimulus-response learning; habit learning; reviewed by Wood & Rünger, 2016). Similarly, repeatedly encountering a neutral stimulus (CS) in the presence of a positive or negative affective stimulus (US) renders the initially neutral stimulus more positive or negative, respectively (e.g., evaluative conditioning; Levey & Martin, 1975; reviewed by Walther et al., 2011). Besides learning based on direct experiences, humans have the capacity to use language as a placeholder for them (e.g., "Don't eat the berries, they will make your stomach hurt"). Language provides an unrestricted combinatory potential. We can comprehend the meaning of a stimulus-response relationship (e.g., berry type and stomach hurt) that we had never directly experienced in that combination before (given that we comprehend the meaning of the individual parts). But do such verbally processed contingencies have a direct effect on subsequent responses? In the present research, we provide further evidence in favor of this assumption. More specifically, we show that repetitive self-instruction of a verbal stimulus-affect contingency in the form of an if-then plan (Gollwitzer, 1999, 2015) influences subsequent evaluative responses towards the stimulus in an valence-congruent direction.

2. Response-compatibility paradigm to assess evaluative reactions

Response-compatibility paradigms are a frequently used method to assess psychological states via behavioral responses (Kornblum et al., 1990); avoiding directly asking for judgments (i.e., self-report). The

☆ The research presented in the manuscript was supported by a Postdoctoral Fellowship to Jenny Roth.
* Correspondence to: T. Martiny-Huenger, UiT The Arctic University of Norway, Department of Psychology, Postboks 6050 Langnes, 9037 Tromsø, Norway.
** Correspondence to: J. Roth, University of Limerick, Limerick V94 T9PX, Ireland.
E-mail addresses: torsten.martiny-huenger@uit.no (T. Martiny-Huenger), jenny.roth@ul.ie (J. Roth).

https://doi.org/10.1016/j.actpsy.2021.103485
Received 11 July 2020; Received in revised form 16 December 2021; Accepted 29 December 2021
Available online 6 January 2022

implicit association test (IAT; Greenwald et al., 1998; see also De Houwer, 2003; Gawronski et al., 2008) is one such paradigm that has been extensively used to assess evaluative responses. Evaluative IATs are designed to elicit a behavioral response bias that indicates evaluative reactions (i.e., the degree of positivity-negativity) towards a target concept. In the test, two categorization tasks are performed using the same two response keys. In a valence-categorization task, participants categorize positive and negative words to a left or right key press. In a target-categorization task, participants use the same left and right key presses to respond to an exemplar of the target category. Valence-based response compatibility between the evaluative reaction towards the target and the positive/negative response keys emerges when performing both tasks in parallel. If the target is assigned to the "positive" key and the target is perceived as positive (i.e., valence-based compatibility), responses are facilitated. In contrast, responses are impaired if the target is assigned to the "negative" key and the target is perceived as positive (i.e., valence-based incompatibility). Consequently, one can estimate the evaluative reaction towards the target concept from the efficiency (i.e., response times and response accuracy) of responses made in the valence-based compatible versus incompatible blocks of the task.

3. Directly experienced stimulus-affect contingencies

Prior research has shown that direct experiences of stimulus-affect contingencies influence responses in evaluative IATs in an evaluatively congruent direction. For example, repeatedly pairing target items like non-words (Mitchell et al., 2003), anime characters (Olson & Fazio, 2001), snack foods (Hollands et al., 2011; Lebens et al., 2011), brands (Gibson, 2008), or particular social groups (French et al., 2013) to affectively negative or positive stimuli resulted in respective changes in the IAT responses. Target concepts repeatedly linked to positive stimuli showed a response bias that indicated more positivity as compared to target concepts repeatedly linked to negative stimuli (and vice versa). This research illustrates that the IAT is sensitive to learning from directly experienced stimulus-affect contingencies (i.e., evaluative conditioning). Based on the literature on directly experienced stimulus-affect contingencies, our present research is concerned with the question whether verbally encountered stimulus-affect contingencies have comparable IAT-measured evaluative consequences.

4. Self-instructed verbal stimulus-affect contingencies

Prior studies have investigated the volatility of IAT responses to different manipulations, including verbal instructions (e.g., Fiedler & Bluemke, 2005; Gregg et al., 2006; Smith et al., 2020; see the General discussion for a more detailed discussion). However, our present research has a more specific focus than mere instructions. We are interested in the effects of self-instructed verbal stimulus-affect plans (see *implementation intentions*; Gollwitzer, 1999, 2015). If-then planning is a self-regulation strategy that links an intended response to a critical cue. The strategy is assumed to create stimulus-response links that are quickly and efficiently activated upon perception of the critical stimulus (Bayer et al., 2009; Martiny-Huenger et al., 2017). Whereas if-then planning research is typically focused on regulating behavior (i.e., by linking an intended behavioral response to a critical stimulus), a small number of prior studies investigated their potential to influence evaluative responses.

For example, Stewart and Payne (2008, Study 3) investigated the effects of verbal stimulus-response planning with the goal of reducing stereotyping. Before measuring participants' evaluative responses towards faces of African Americans with an IAT, they were instructed to silently say to themself: "[…] I definitely want to respond to the Black face by thinking 'good.'" These instructions included the verbal stimulus-response contingency: "Black face" and thinking "good". Linking the positive response to the targets resulted in a reduced negativity-indicating response bias compared to a control condition that linked an evaluatively neutral concept to the targets (see Lai et al., 2014, 2016 for direct replications of this effect).

Similarly, Hofmann et al. (2010, Study 2) reduced positive evaluative responses towards chocolate. Before measuring participants' evaluative responses towards chocolate with an IAT, they were instructed to formulate a specific plan to avoid chocolate. The example given to participants as an orientation was "If my friend offers me chocolate during the film, I will say 'no thanks' and concentrate on the film." Assuming that participants followed these instructions, they processed a verbal stimulus-response contingency that linked "chocolate" to an (evaluatively negative) avoidance response. The results showed a reduced positivity-indicating response bias after participants rehearsed the verbal stimulus-response contingency as compared to a control condition.

Our present research is conceptually similar to these previous studies. However, in contrast to the categorical valence content ("good") used by Stewart and Payne (2008) and the behavioral responses that implied negativity (saying "no thanks"), we linked verbal descriptions of a clearly affective concept ("disgusting fat" vs. "delicious sweets") or clearly affective response ("disgusting" vs. "delicious") to unhealthy food items. This aligned the verbal descriptions of the affective content used in our present studies to previous studies investigating the consequences of direct experiences on evaluative responses measured by the IAT (e.g., aversive obese bodily images; Hollands et al., 2011). We could thereby more clearly investigate whether encountering a verbal stimulus-affect contingency has similar evaluative consequences like the previously listed studies investigating stimulus-affect contingencies that were directly experienced (French et al., 2013; Gibson, 2008; Hollands et al., 2011; Lebens et al., 2011).

Furthermore, on a methodological level, the study reported by Hofmann et al. (2010, Study 2) is not informative about whether the observed reduction in positivity towards chocolate was a consequence of the verbal stimulus-response contingency or a consequence of highlighting the goal of eating less chocolate – a goal that was not present in the control condition. A similar conceptual ambiguity can be found in a recent meta-analysis (Forscher et al., 2019), where the authors categorized the procedure used by Stewart and Payne (2008) as a goal-induced effect. Thus, empirically and theoretically, it is not clear whether the previously observed effects are a consequence of (or attributed to) mechanisms of goal setting or merely to encountering a stimulus-response contingency (in a verbal format).

In the present research, we avoided such ambiguity between goal setting (e.g., "wanting to eat less chocolate") and the processing of a verbal stimulus-affect contingency by providing all experimental groups with the same goal-related information (i.e., "I want to eat fewer cupcakes") and varying only the valence (disgusting vs. delicious) of the verbal stimulus-affect contingency. Consequently, any observed effects in our present studies are more likely to be attributed to the verbal stimulus-affect contingency than other motivational aspects.

5. The present research

We tested the consequences of processing verbal if(stimulus)-then (affect) plans on evaluative responses in three studies. Whereas all participants were asked to commit to the same goal (eating fewer cupcakes), we presented participants with either a positive versus negative stimulus-affect contingency in the form of a verbal if-then plan (negative plan: "If I see a cupcake, then I will think 'disgusting'" versus positive plan: "If I see a cupcake, then I will think 'delicious'"). Subsequently, we assessed evaluative responses towards cupcakes with a (ST-)IAT (i.e., response-compatibility paradigm) and self-reported ratings.

Regarding the IAT responses, we hypothesized that participants who linked a negative affective concept (e.g., "disgusting") to cupcakes would show IAT responses indicating less positivity (/more negativity) towards cupcakes compared to participants who linked a positive

Table 1
Descriptive statistics of the participant samples and central study differences.

	Study A	Study 1	Study 2
Descriptive statistics			
N (female)	80 (64)	117 (98)	126 (107)
Age *M* (*SD*)	20.4 (5.6)	20.1 (2.5)	19.7 (2.3)
Age min., max.	18, 60	18, 38	18, 34
Central study differences			
Test paradigm	IAT	ST-IAT	ST-IAT
Key-assignment switch	Valence-key switch	Target-key switch	Target-key switch
Valence-concept formulation	"Delicious sweets"/"disgusting fat"	"Delicious sweets"/"disgusting fat"	"Delicious"/"disgusting"

affective concept (e.g., "delicious") to cupcakes. More specifically, in IAT blocks with the targets assigned to the "positive" response key, participants in the positive plan-valence condition should show facilitated responses (quicker and more accurate) than in blocks with the targets assigned to the "negative" response key. Compared to the positive plan-valence condition, this difference should be smaller (or reversed) in the negative plan-valence condition.

Besides analyzing the response times and response errors in the different target-response assignment blocks of the IAT, the same effect should be observed in the IAT's D-score. The D-score is a single score that integrates response times and accuracy in the valence-based compatible and the incompatible parts (Greenwald et al., 2003). Participants in the negative plan-valence condition should show a lower D-score value (i.e., indicating more negativity) than participants in the positive plan valence condition. Finally, regarding the self-report ratings, we expected participants in the negative plan-valence condition to report lower self-reported liking of cupcakes than participants in the positive plan-valence condition. Importantly, we expected to find these effects despite all participants receiving the same evaluatively negative goal-related information (i.e., committing to eating fewer cupcakes).

Because of the high degree of similarity between the studies, we present all methods and results in a single method and result section below. In this final section of the introduction, we summarize the central differences between the studies regarding the IAT variant used, the verbal formulation of the affective response, and the method for establishing the IAT's target-response assignment blocks. First, in Study A (the preliminary study), we used a standard IAT in which the target-categorization task was based on a cupcake (target) and house (control) categorization. In Studies 1 and 2, we used a single-target IAT (ST-IAT) based on the same response-compatibility principle but without requiring a second (house) response category (Bluemke & Friese, 2008). Second, concerning the verbal affective response, in Studies A and 1, we linked verbally-expressed concepts ("disgusting fat" versus "delicious sweets") to the target category of cupcakes. In Study 2, we replaced these concepts with a verbal expression of the affective response itself (i.e., "disgusting" and "delicious") without referring to objects (fat and sweets).

Finally, there is a methodological difference in establishing the different compatibility blocks that made us distinguish the first study (Study A) from the latter two studies (Studies 1 & 2). Assessing valence-based response compatibility requires two blocks with different target-response assignments: one in which the target response is made with the "positive" key and one in which the target response is made with the "negative" key. In Study A, we kept the target-response key (left key) constant over the two blocks and switched the valence assignment between the left and right keys to establish the two blocks. In one block, the left key was assigned to positivity; the same key was assigned to negativity in the second block. This choice might have undermined establishing a strong link between the valence and the respective key in the second block because the participants were required to relearn the key-valence assignment and forget the old one. As we did not find evidence for our predictions in Study A, we held the valence-key assignment constant over both blocks in Studies 1 and 2. Instead, the target-response key assignment was switched between the two blocks. Following this change, in Studies 1 and 2, we found evidence for the hypothesized consequences of verbal stimulus-affect contingencies on the IAT-measured evaluative responses.

6. Methods

6.1. Participants and design

Participants were recruited through a participant pool system at University of California Davis (USA) in return for course credit. Table 1 shows descriptive participant statistics for all three studies. No power analyses were conducted prior to the studies. Instead, the sample size was set to approximate 50 participants per between-participant condition following related prior laboratory studies (Stewart & Payne, 2008). No analyses were conducted before the full reported sample was collected. All three studies followed a 2 by 2 design with the between-participant factor *plan valence* (positive [delicious] vs. negative [disgusting]) and the within-participant factor *target-response assignment* (positive key vs. negative key). The dependent variables were response times and response errors in the IATs (Study A: IAT and Studies 1 & 2: ST-IAT). In addition to the detailed analysis of response times and response errors in the critical IAT blocks, we calculated the typically used IAT D-score as a dependent variable that combines response times, errors, and the critical IAT blocks into a single score. Consequently, the IAT D-score analysis only followed a one-factorial design (plan valence: positive vs. negative). The self-report attitude ratings also followed a 2 by 2 factorial design but with the between-participant factor *plan valence* (positive [delicious] vs. negative [disgusting]) and the within-participant factor *stimulus type* (target [cupcakes] vs. control [houses]). The dependent variable of the self-report measurement were attribute ratings and a thermometer scale rating of the target (cupcakes) and control (houses) stimuli.

6.2. Procedure and materials

Participants were informed that the study was about food choices and health. In line with typical procedures in if-then planning investigations, each study started with the goal setting, followed by the experimental manipulation of the verbal stimulus-affect contingency (i.e., plan-valence condition). After the critical experimental manipulation and before assessing the dependent variables, participants were asked about their level of goal and plan commitment. Finally, participants completed the response-compatibility paradigm (IAT or ST-IAT) and the self-reported liking ratings. A demographic questionnaire, including age and gender, was presented at the end of the study. The following sections provide the details of the procedure in the order of presentation.

6.2.1. Goal setting

All participants received the same goal-related instructions to ensure that the results can be attributed to the rehearsed verbal plans and not differences in goals. We informed the participants that considering one's health is a common task in everyday life that must be implemented

continuously (e.g., not eating a cupcake during a coffee break). They read that the study was about testing simple strategies to help them avoid unhealthy snacks (e.g., cupcakes). Furthermore, we added an explanation that made the negative and the positive plan a plausible strategy to reach that goal. Finally, we asked them to commit to the goal to avoid eating cupcakes ("My Goal: I want to avoid eating cupcakes!").

6.2.2. Plan-valence condition

After the goal instruction, we provided all participants with the verbal plans (i.e., verbal stimulus-affect contingency). The plans linked either a positive or a negative concept to the target stimuli. In Studies A and 1, the plans were "*Whenever I see a cupcake, then I will think of delicious sweets!*" (positive plan valence) versus "*Whenever I see a cupcake, then I will think of disgusting fat!*" (negative plan valence). In Study 2, the same plans were used, but the valence concept did not include the objects ("sweets", "fat") but only the affective response ("delicious" vs. "disgusting").

The memorization and repetition of the respective plan was assured by a three-step procedure: First, the plan was presented on the computer screen, and we asked the participants to memorize it and to take 1 min to rehearse it silently. Second, the plan was removed from the computer screen and we asked the participants to write it on a sheet of paper (paper and pen were available on the table). Third, we presented the plan again on the computer screen and asked the participants again to take a minute to rehearse it. This procedure assured that the stimulus-affect contingency was "encountered" at least twice visually as words on the computer screen, once in self-produced handwriting, and an unspecified number of times as silent verbal articulation during the two 1-minute periods of rehearsal.

6.2.3. Goal commitment

We assessed participants' goal and plan commitment with six items (e.g., "I am strongly committed to pursuing the goal to avoid cupcakes", 1 = *strongly disagree*, 7 = *strongly agree*; Cronbach's alpha for Study A, 1, and 2, 0.74, 0.72, and 0.82, respectively). We combined all items within each study to one commitment score. Two-sided *t*-tests indicated that there is a significant difference in the commitment score of Study A between goal commitment in the positive plan-valence condition ($M = 4.7$; $SD = 1.0$) and negative plan-valence condition ($M = 5.2$; $SD = 1.1$), $t(75.9) = 2.11$, $p = .038$, CI [0.03, 0.96]. No evidence for such a difference was found in Study 1 (positive plan, $M = 4.8$; $SD = 0.9$; negative plan, $M = 4.8$; $SD = 1.1$), $t(108.2) < 1$, $p = .836$, CI [−0.33, 0.41] or Study 2 (positive plan, $M = 4.9$, $SD = 1.2$; negative plan, $M = 5.0$, $SD = 1.3$), $t(122.1) < 1$, $p = .585$, CI [−0.31, 0.54]. Although there is no evidence of different levels of goal commitment following the goal commitment and planning procedure in the central Studies 1 and 2, we included goal-commitment scores as a covariate in all analyses of our central hypothesis.

6.2.4. Response-compatibility paradigms (IATs)

Participants read the instructions for the task ahead of each block, and the category labels were displayed throughout the task in the upper-left and upper-right corners of the screen. Stimuli were presented in random order with an intertrial interval of 150 ms. A red "X" was used to denote incorrect responses. It remained on the screen until the correct response was made.

6.2.4.1. Standard IAT (Study A). The implemented IAT consisted of three practice blocks and two critical blocks (i.e., target-response assignment factor). In the first practice block, participants practiced the target-categorization task by categorizing three different pictures of cupcakes and three different pictures of houses to the categories "cupcake" and "house", respectively (6 trials). In the second block, participants practiced the valence-categorization task by categorizing three positive attributes (appealing, pleasant, and tempting) and three negative attributes (repulsive, revolting, and unpleasant) to the respective positive and negative category (6 trials). In the first critical block, participants executed both categorization tasks in parallel. Importantly, the cupcake category was assigned to the same response key as the positive-word category (A key), and the house category was assigned to the same response key as the negative word category (Numpad 5, 72 trials). Then, the key assignment of the valence categories was switched and participants practiced the reversed categorization of the valence attributes (6 trials). Finally, in the second critical block, participants performed both categorization tasks in parallel again. By switching the valence-key assignment, the cupcake category now overlapped with the negative-word category (A key), and the house category overlapped with the positive-word category (Numpad 5, 72 trials). The two critical blocks contained a short break after the first 36 trials.

6.2.4.2. Single-target IAT (Studies 1 & 2). The ST-IAT consisted of one practice block and two critical blocks (i.e., target-response assignment factor). Participants started with practicing the valence-categorization task (12 trials), in which they categorized three positive attributes and three negative attributes (see Study A) to the positive-word category (A key) and negative-word category (Numpad 5; the attribute "tempting" was replaced with "yummy" in Study 2). In the following two critical blocks, participants executed both the valence-categorization task and the target-categorization task with the same valence-key assignment throughout both blocks. In contrast to the standard IAT used in Study A, the target-categorization task included only the critical target stimuli (cupcakes). Whenever a picture of a cupcake appeared, participants pressed the assigned target-response key. In the first critical block, the target-response key was the same as the positive-valence key (A key, 72 trials). For the second critical block, the target-response key was reassigned to the negative-valence key (Numpad 5, 72 trials). Both critical blocks included a short break after 36 trials. Depending on the block, cupcake stimuli, positive attributes, and negative attributes occurred in a ratio of 18:18:36 trials per critical block (e.g., 18 cupcakes trials, 18 positive-word trials, and 36 negative-word trials). This led to an equal proportion of left-hand and right-hand responses in each of the critical blocks (see Bluemke & Friese, 2008).

6.2.5. Self-report ratings

Participants rated the extent to which different attributes (positive: appealing, delicious,[1,2] pleasant, tempting,[3] and yummy[3]; negative: repulsive, revolting, unpleasant, and disgusting[1]) describe the target and control stimuli (e.g., "Cupcakes are appealing", 1 = *strongly disagree*, 7 = *strongly agree*). Ratings of cupcakes were assessed before the ratings of houses, but the order of attribute presentation within these categories was randomized. Negative-valence attributes were reverse scored for the analysis so that higher scores indicate more positive ratings. Additionally, we assessed global evaluations of cupcakes and houses using a thermometer scale ("On a feeling thermometer from 0 to 100, how positive do you feel about [cupcakes/houses]?"). In Study 2, we used a slightly different wording and replaced "positive" with "attracted." One missing response to the house-thermometer scale in Studies A and 2 was replaced by the median calculated from the remaining responses. Internal consistency for the self-reported ratings was high. Cronbach's alpha for the Studies A, 1, and 2 for cupcake items was 0.92 (9 items), 0.92 (7 items), and 0.93 (7 items), respectively and for house items 0.81 (7 items), 0.86 (6 items), and 0.77 (6 items).

[1] "Delicious" and "disgusting" were not used to rate houses.

[2] "Delicious" was omitted in Studies 1 and 2 to avoid an obvious overlap with the word used in the verbal target-valence contingency.

[3] "Tempting" was replaced with "yummy" in Study 2 as prior participants indicated that "tempting" was not perceived as an unambiguously positive attribute.

6.3. Data preparation and analyses

We analyzed only responses from the critical (ST-)IAT blocks with parallel target-categorization and valence-categorization responses (72 trials per block). Descriptive analysis of mean response times (Study A: $M = 740$ ms, $SD = 147$ ms; Study 1: $M = 668$ ms, $SD = 132$ ms; Study 2: $M = 632$ ms, $SD = 97$ ms) and mean response error percentage (Study A: $M = 9.0\%$, $SD = 5.1\%$; Study 1: $M = 9.6\%$, $SD = 6.1\%$; Study 2: $M = 7.9\%$, $SD = 6.1\%$) per participant indicated no conspicuous mean participant data. In line with IAT analysis criteria (Greenwald et al., 2003), we removed the complete data of participants with more than 10% responses below 300 ms (Study A: none; Study 1: 1 participant [0.9%]; Study 2: 3 participants [2.4%]). From the resulting data, we removed responses below 200 ms (fast guesses; Study A: 0.1%; Study 1: 0.5%; Study 2: 0.1%). Finally, for response time analyses, error responses (Study A: 9.0%; Study 1: 9.4%; Study 2: 7.3%) and responses beyond three standard deviations (*SD*s) below or above the mean calculated per participant and response-key overlap block (Study A: 1.7%; Study 1: 2.1%; Study 2: 2.1%) were removed.

R (R Core Team, 2014) was used for all analyses. Linear mixed-model analyses (lme4 package; Bates et al., 2015) were used to analyze the response times, response errors (binomial), and self-report ratings. Response time analyzes were performed on the log-transformed values. Besides the relevant fixed effects for each respective analysis, we specified the IAT-block by participant slope as random effect. We obtained *p*-values from the stats package's anova function (Chambers & Hastie, 1992) for response times and self-report data and the car package's Anova function (Fox & Weisberg, 2011) for the binomial response error data. We also analyzed the (ST-)IAT by calculating the D-score (improved algorithm; Greenwald et al., 2003). The D-score reflects the response times and response errors (i.e., error responses are replaced by the block mean plus a 600 ms time penalty) in the two critical blocks as a single value. More positive values indicate more positivity of the targets. The D-scores calculated for each participant were analyzed using the ezAnova function from the ez-package (Lawrence, 2016).

7. Results and discussion

7.1. Central hypothesis (IAT)

Our central hypothesis is reflected by a 2-way interaction effect between plan-valence condition and target-response assignment. The response time and error difference between the negative and positive target-response assignment blocks (negative minus positive) should be smaller in the negative plan-valence condition than in the positive plan-valence condition. The results of this 2-way interaction effect for all studies are presented in Table 2. Fig. 1 illustrates the direction of the interaction effect in the respective plots for Study 1 and 2 (see the Supplemental material for the Study A plot). Fig. 2 illustrates the confidence intervals for the central 2-way interaction effect. There is no indication of the predicted interaction effect in Study A. As described in the introduction and methods section, we updated the IAT procedure for Studies 1 and 2 (i.e., SC-IAT and implementing the two IAT blocks by switching the key assignment of the target-categorization task instead of using the valence-categorization task). Following this change, for Studies 1 and 2, all indicators (response times, response errors, and D-score) show the predicted 2-way interaction effects between the plan-valence condition and the target-response assignment. The only exception is the response-time analysis of Study 1. However, as illustrated by the plot (Fig. 1A) and the confidence intervals (Fig. 2), also this non-significant effect shows a pattern in the predicted direction.

Table 2
Response time, response error, and D-score analyses for the central hypothesis in Studies A, 1, and 2.

	F/χ^2	*p*	Low CI	High CI
Study A (IAT, valence-key reversal)				
Response times	0.40	.530	−0.021	0.011
Response errors	0.80	.371	−0.051	0.136
D-score	0.08	.778	−0.063	0.095
Study 1 (ST-IAT, target-key reversal)				
Response times	1.04	.310	−0.004	0.013
Response errors	9.74	.002	0.033	0.142
D-score	3.90	.051	0.026	0.155
Study 2 (ST-IAT, target-key reversal)				
Response times	4.21	.042	0.0004	0.016
Response errors	4.58	.032	0.004	0.122
D-score	4.36	.039	0.032	0.159

Note. Statistical values for response times and response errors represent the interaction effect between the plan-valence condition and target-response assignment block. The D-score values represent the main effect of the plan-valence condition as the target-response assignment blocks are integrated in the D-score calculation.

The response time plots and response error plots (Fig. 1) confirm that the interaction effect patterns align with our predictions. In Studies 1 and 2, response times are lower and response errors are fewer in the positive target-response assignment blocks compared to the negative target-response assignment blocks (i.e., indicating positivity). Importantly, these differences are smaller (i.e., indicating less positivity/more negativity) in the negative plan-valence condition than in the positive plan-valence condition. Furthermore, we observed these effects while controlling for the level of goal commitment. No analysis indicated that goal commitment moderated the central 2-way interaction effect (all *p*'s $> .180$; see also the Testing the relevance of goal commitment section below).

In sum, in Studies 1 and 2, we found evidence that participants receiving a plan that linked a negative affective concept to cupcakes showed a response bias in the IAT that indicated less positivity (more negativity) than participants receiving a plan that linked a positive affective concept to cupcakes. In Study A, which we conducted prior to Studies 1 and 2, we did not find this effect. Besides the use of a traditional evaluative IAT in Study A and evaluative ST-IATs in Studies 1 and 2, a central difference between the studies was the approach on how to establish the two target-response assignment blocks. In Study A, this was achieved by switching the valence-response key assignment between the two IAT blocks and keeping the target-response assignment constant. In Studies 1 and 2, this was achieved by switching the target-response key and keeping the valence assignment constant. Compatibility effects between the evaluative responses towards the targets and the valence of the response keys depend on a strong key-valence association. The procedure to switch this valence assignment from one block to the next in Study A may have undermined such a strong key-valence association in the second block. Consequently, the required strong key-valence association may not have been established successfully in the second block. In Studies 1 and 2, the key-valence assignment was kept constant to guarantee a strong key-valence association also in the second block. With this procedure, we found the predicted effects. Independent additional tests are required to verify this post-hoc reasoning.

7.2. Testing the relevance of goal commitment

To have a robust test for any moderation effect by goal commitment, we combined the data of Studies 1 and 2. Based on our prior reasoning, these two studies are methodologically more solid than Study A, and the increased power resulting from combining the data should decrease the likelihood of finding spurious effects. The results of all relevant or significant results regarding goal commitment are listed in Table 3. To anticipate the conclusion, we find no evidence that our central results are undermined by goal commitment as a confounding variable.

7.2.1. Response errors

Importantly, we found evidence for the predicted plan valence by

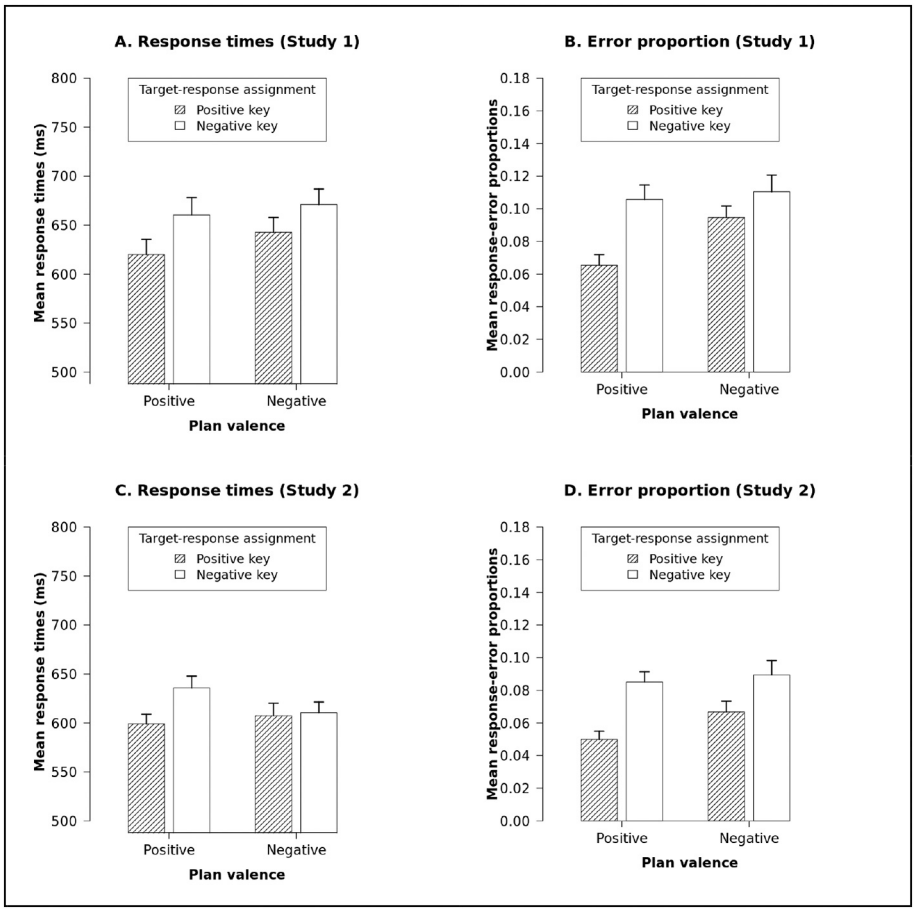


Fig. 1. Mean response times (left) and accuracy (right) for the plan-valence by target-response assignment factors of Studies 1 and 2. Note. Whiskers represent one standard error of the mean (SEM).

target-response assignment interaction effect ($\chi^2 = 14.50, p < .001$), and this predicted 2-way interaction effect is not further qualified by an interaction with goal commitment ($\chi^2 = 0.42, p = .518$). The only other noteworthy effect is a main effect of goal commitment. The higher the goal-commitment score of a participant, the lower the number of response errors.

7.2.2. Response times

Similar to the response errors, we found evidence for the predicted plan valence by target-response assignment interaction effect ($\chi^2 = 4.44$, $p = .036$), and this predicted 2-way interaction effect was not further qualified by an interaction with goal commitment ($\chi^2 = 1.04, p = .310$). There was no evidence for a main effect of the goal-commitment score on response times.

7.3. Self-report ratings

To test for effects of the verbal stimulus-affect contingencies on self-reported liking ratings, we conducted linear mixed-model analyses on the mean standardized ratings for each item (i.e., attributes and thermometer) with the fixed effects item valence (positive vs. negative attribute), plan valence condition (positive vs. negative), stimulus type (target vs. control stimuli), and the interaction term between plan valence and stimulus type. Intercepts for participant ID and rated item were entered as random effects. In all three studies, the analysis revealed significant plan valence by stimulus type interaction effects (Study A: $\chi^2 = 40.89, p < .001$; Study 1: $\chi^2 = 6.37, p = .012$; Study 2: $\chi^2 = 39.14, p < .001$). The result patterns are plotted in Fig. 3, and descriptive statistics are presented in detail in the Supplementary material (Table 4). The plots indicate that the interaction effect is a result of the predicted effect of the plan-valence condition. Target stimuli (cupcakes) were rated more negative in the negative plan-valence condition than in the positive plan-valence condition. In contrast, no such difference was observed in the ratings of the control stimuli (houses). Thus, the analysis provides evidence that the experimental manipulation affected the target stimuli – but not the control stimuli – in the predicted direction.

8. General discussion

The present studies provide further evidence that verbally processed stimulus-affect contingencies (i.e., if-then plans) influence subsequent (evaluative) responses towards the respective targets. We observed these consequences in a valence-based response-compatibility paradigm (ST-IAT) and self-reported ratings. The response patterns of Studies 1 and 2 (and Study A for self-reports) indicate that the target stimuli were perceived as less positive (more negative) after participants committed themselves to think "disgusting" (as compared to "delicious") whenever they encountered the target.

Before continuing the discussion, we want to evaluate our results regarding some criticism raised about the validity of the IAT-measured response bias (e.g., Fiedler et al., 2006; Kinoshita and Peek-O'Leary, 2005; Rothermund & Wentura, 2004). There is evidence that under certain circumstances, valence-based compatibility effects can be confounded with familiarity (Kinoshita and Peek-O'Leary, 2005), saliency (Rothermund & Wentura, 2004), or easily perceived category similarities (Bading et al., 2020). Whereas we can exclude some non-valence-based alternative explanations for our present research (e.g., the single-target IAT variant does not include differently familiar target

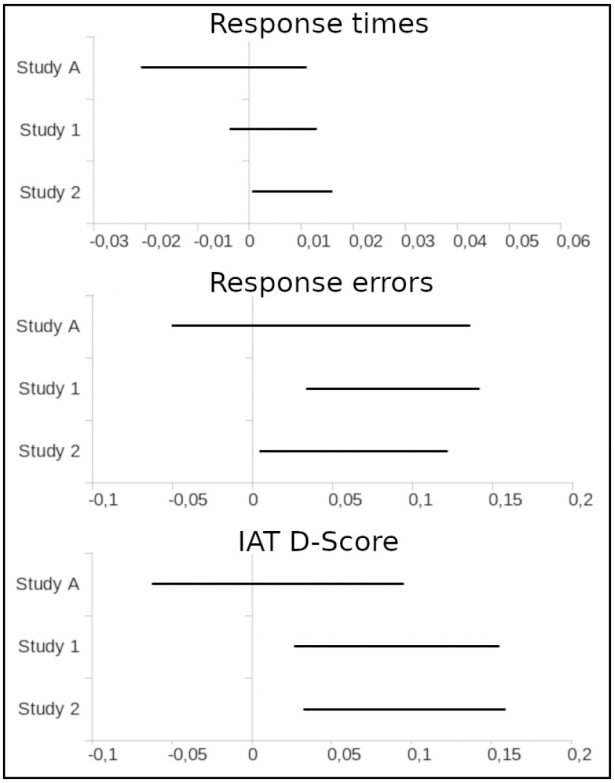


Fig. 2. Confidence intervals for the central plan-valence by target-response assignment interaction for Studies A, 1, and 2.
Note. Because the IAT D-score calculation combines the two IAT compatibility blocks, the confidence intervals only conceptually reflect the interaction pattern but are actually the main effect of the plan-valence condition.

Table 3
Goal-commitment related main effects and interaction effects of the combined data of Studies 1 and 2.

Factors and interaction terms	χ^2/F	p	CI
Response errors			
Goal commitment	6.09	.014	−0.210, −0.029
Target-response assignment × goal commitment	3.22	.073	−0.001, 0.080
Study × plan valence × goal commitment	5.76	.016	0.025, 0.206
Plan valence × target-response assignment × goal commitment	0.42	.518	−0.050, 0.030
Response times			
Goal commitment	0.61	.433	−0.023, 0.010
Target-response assignment × goal commitment	5.66	.018	0.001, 0.013
Study × target-response assignment × goal commitment	2.78	.097	−0.001, 0.011
Plan valence × target-response assignment × goal commitment	1.04	.310	−0.009, 0.003

Note. Confidence intervals (CI) represent the 2.5% and 97.5% borders (i.e., 95% CI).

and comparison concepts), we cannot completely exclude other alternatives (e.g., saliency-based and similarity-based recoding of the instructions). However, despite these potential confounding variables, IAT results correlate with self-reported valence measures (reviewed by Greenwald et al., 2009; Hofmann et al., 2005; see also Greenwald et al., 2005; correlation in the present Studies 1 and 2 combined ($N = 239$): $r = 0.175$, $p = .007$, CI (0.049–0.295)). Consequently, even if contributions from alternative processes cannot be excluded completely, valence-based compatibility effects are likely to contribute to the observed effects in the present studies. Furthermore, the widespread prior use of the IAT allows comparing our present results with this previous research.

8.1. Relation to directly experienced stimulus-affect contingencies and other instruction-based evaluative consequences

Our present results align with prior research assessing evaluative changes induced by directly experienced stimulus-affect contingencies (French et al., 2013; Gibson, 2008; Hollands et al., 2011; Lebens et al., 2011). For example, Hollands et al. (2011) repeatedly paired snack foods with aversively obese body images and found reduced IAT-measured preferences for the targeted snacks. We did not present any negative or aversive images in our present research. Instead, participants merely linked the words “disgusting fat” or “disgusting” to cupcakes. Conceptually in line with the results from Hollands et al., our present procedure of linking a negative affective word to cupcakes reduced IAT-measured preferences for cupcakes compared to linking a positive word to them.

Our present research has similarities with research on instructed evaluative conditioning (e.g., De Houwer, 2006; Kurdi & Banaji, 2017; Smith et al., 2020; see also Gast & De Houwer, 2013 for verbal extinction instructions). In general, this research is concerned with the evaluative consequences of instructing that a specific stimulus (CS) will later be presented together with a positive or negative concept or image (US). The target stimuli (CS) are never paired with the positive or negative items (US; other than in the instructions). Still, typical results align with research on directly experienced stimulus-valence contingencies. Evaluative reactions become more positive (more negative) if the target stimulus is instructed to be later presented together with a positive (negative) item (e.g., De Houwer, 2006).

There is a central difference between instructed evaluative conditioning and our present approach. Instructed evaluative-conditioning studies inform about upcoming stimulus-valence pairings. Thus, they create an expectancy that a specific stimulus will later co-occur with a positive or negative item. The planning procedure in our present research does not establish such expectancies. Instead, it induces a commitment to producing a specific cognitive response (e.g., think ‘disgusting’; US) upon encountering the target stimulus. Such a commitment to a verbal if-then plan is assumed to establish a link between the situation (e.g., cupcake) and response (e.g., thinking ‘disgusting’; reviewed by Gollwitzer & Sheeran, 2006). In line with this idea, we observed that evaluative reactions towards the targets were in line with the novel affective information (i.e., less positive evaluation if the link concept was negative than positive; see the next section for alternative explanations).

Although the if-then planning procedure (our present studies; Hofmann et al., 2010, Study 2; Stewart & Payne, 2008, Study 3) and instructed evaluative conditioning (De Houwer, 2006) appear to produce similar evaluative outcomes (but see Mattavelli et al., 2021), future research may gain novel insights into the underlying mechanisms by focusing on differences between the procedures. For example, instructed evaluative conditioning informs about an upcoming co-occurrence of CS and US that never happens. Thus, there may be a conflict between expectation and observation when encountering the CS. This is not the case for the if-then planning procedure. Participants do not expect to encounter the US in the presence of the CS; they planned to “produce” the US themselves. What conclusions can we draw from observing similar evaluative consequences despite the differences in the procedures? Does the planning procedure unintendedly produce similar expectations as instructed evaluative conditioning? Alternatively, does instructed evaluative conditioning unintendedly make participants “self-produce” the expected but missing US? Disentangling these possibilities in future research may help us to better understand the mechanisms of how verbal information influences subsequent responses.

Finally, recent studies on instructed evaluative conditioning have

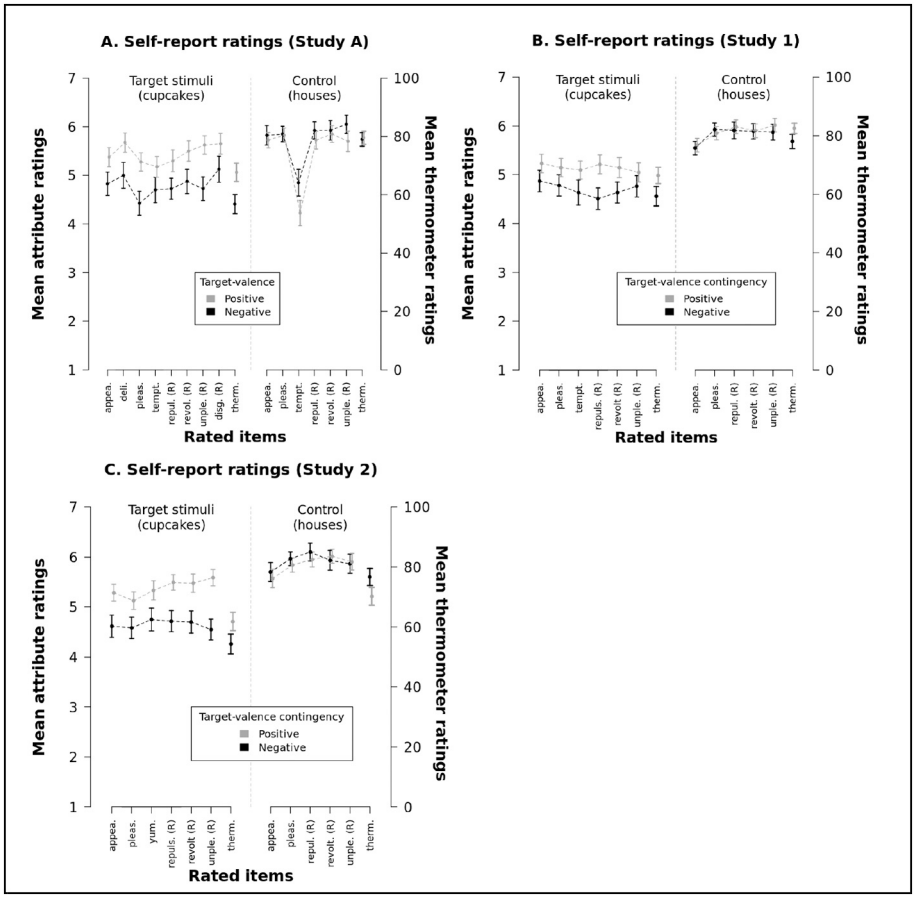


Fig. 3. Mean self-report ratings for plan valence by stimulus type in Studies A, 1, and 2.
Note. Whiskers represent one standard error of the mean (SEM) above and below the mean. The second y-axis refers to the respective single data point on the right-hand side of each graph.

reported unexpected IAT effects (De Houwer et al., 2019; Mattavelli et al., 2021). It appears that minor changes in the instructions can lead to different results. The potential of finding surprising effects based on small changes in the introduction highlights our present research's value of conceptually replicating prior studies using the if-then planning procedure. In that regard, our present results consolidate prior research that assessed the evaluative consequences of verbal if-then plans (Hofmann et al., 2010, Study 2; Stewart & Payne, 2008, Study 3) and, to some degree, indicates that such effects can be observed with different formulations. Our studies are most similar to the study done by Hofmann et al. (2010, Study 2), with a common focus on changing evaluative responses towards unhealthy sweet snacks. We provide a conceptual replication with clearly formulated affective responses in the then-part of the plan. Furthermore, we add to this research by separating mechanisms of goal setting and processing the stimulus-affect (i.e., if-then) contingency. In contrast to the study reported by Hofmann et al. (2010, Study 2) where the goal of eating less chocolate was not present in the control condition, we provided the same goal ("I want to eat fewer cupcakes") to all participants and only manipulated the valence of the concept linked to cupcakes. Our results can be more unambiguously attributed to the differently valenced verbal stimulus-affect contingency than to mechanisms of goal setting (e.g., Fishbach et al., 2004).

8.2. *Different perspectives on the effects of verbal information on behavior and affect*

If-then planning effects are traditionally assumed to result from associative links between the verbally linked if- and then-part (e.g., Bayer et al., 2009; reviewed by Gollwitzer & Sheeran, 2006). In contrast, researchers investigating instructed evaluative conditioning often highlight propositional beliefs to explain the observed effects (e.g., De Houwer, 2006; Smith et al., 2020). We believe that it is essential to highlight that linking the former or latter theory to the effects reported in the present research is rather a theoretical choice than an empirical one.

We cannot interpret empirical evidence independently from theoretical assumptions. For example, De Houwer (2006; see also Smith et al., 2020; Schmidt et al., 2016) presents research on instructed evaluative conditioning and concludes that propositional beliefs influence the IAT. The author, however, acknowledges that this conclusion depends on the assumption that processing verbal stimulus-response/affect contingencies does not result in associative learning as the stimulus (CS) and "response" (US) never were paired (De Houwer, 2006, p. 185; see also Smith et al., 2020, p. 30 for the inclusion of a similar assumption). In our present studies, the stimulus (cupcakes) and affective response (experience of disgust) were never directly paired. Thus, based on the assumption made by De Houwer (2006) that associative learning requires directly experienced pairings, our present results cannot be explained by associative learning and are more likely driven by propositional beliefs.

The validity of this conclusion, however, depends on the validity of the pre-assumption that associations only form from direct experiences. Simulation theories of cognition (i.e., grounded cognition; Barsalou, 1999; Barsalou, 2016; Hesslow, 2012) provide a basis to doubt the validity of the assumption. The cluster of theories and the related empirical evidence indicates that thought (e.g., language comprehension) operates on simulated experiences located in brain areas involved in processing the direct experiences. Thus, if processing verbal

representations of a situation and processing of the direct experiences of the situation are mediated by the same brain areas and overlapping activity patterns, both can be expected to result in overlapping associative-learning outcomes (see Martiny-Huenger et al., 2015; Martiny-Huenger et al., 2017 for a more extended elaboration on this argument). As this associative-learning perspective – from a connectionist associative learning perspective (e.g., Smith & Conrey, 2007) – is in line with previous evidence (e.g., De Houwer, 2006), conclusions about associative versus propositional mechanisms cannot easily be drawn by the research demonstrating the effects (e.g., De Houwer, 2006; Smith et al., 2020; our present studies). Instead, research on the (in) validity of central pre-assumptions (e.g., how is verbal information processed and comprehended; see Barsalou, 2016; Fischer & Zwaan, 2008) are required to settle the argument.

9. Conclusion

Our present research contributes to the evidence that verbal stimulus-affect contingencies influence subsequent behavioral responses that, in the case of the IAT, are typically interpreted as reflecting valence differences (degrees of positivity-negativity). Furthermore, in contrast to research on instructed evaluative conditioning (e.g., De Houwer, 2006; Kurdi & Banaji, 2017), we found these effects in a context where participants were not informed about actual co-occurrence of the CS and US in the future (i.e., creating an expectancy), but by a self-commitment to "produce" the US in thought any time the CS is encountered. Regarding previous research using similar planning strategies (Hofmann et al., 2010, Study 2; Stewart & Payne, 2008, Study 3), our present research replicates these effects with more explicit affective concepts as US and separates them from more motivational mechanisms.

In general, the effects of indirect (e.g., verbal) information on responses are intriguing as they provide explanations for the apparent complexity of human behavior that is hard to explain solely as a consequence of direct experiences. The unrestricted freedom to verbally combine stimuli and behavioral and affective responses can serve as a basis for acquiring novel stimulus-response/affect links that guide subsequent responses. Last but not least, this flexibility of not being restricted to actual experiences highlights the potential to use this verbal flexibility in self-instructions to influence one's own (evaluative) responses towards certain items (e.g., reducing positivity of unhealthy snacks) in a way that aligns the evaluative response to a specific goal (e.g., eating less unhealthy food).

Declaration of competing interest

No conflicts to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2021.103485.

References

Bading, K., Stahl, C., & Rothermund, K. (2020). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion, 34*, 128–143. https://doi.org/10.1080/02699931.2019.1604322

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660. https://doi.org/10.1017/S0140525X99232141

Barsalou, W. L. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review, 23*, 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4*.

Bayer, U. C., Achtziger, A., Gollwitzer, P. M., & Moskowitz, G. B. (2009). Responding to subliminal cues: Do if-then plans facilitate action preparation and initiation without conscious intent? *Social Cognition, 27*, 183–201. https://doi.org/10.1521/soco.2009.27.2.183

Bluemke, M., & Friese, M. (2008). Reliability and validity of the single-target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977–997. https://doi.org/10.1002/ejsp.487

Chambers, J. M., & Hastie, T. J. (1992). *Statistical models in S*. Wadsworth & Brooks/Cole.

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch, & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Routledge Taylor & Francis Group.

De Houwer, J. (2006). Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37*, 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J., Mattavelli, S., & Van Dessel, P. (2019). Dissociations between learning phenomena do not necessitate multiple learning processes: Mere instructions about upcoming stimulus presentations differentially influence liking and expectancy. *Journal of Cognition, 2*, 7. https://doi.org/10.5334/joc.59

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology, 27*, 307–316. https://doi.org/10.1207/s15324834basp2704_3

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the implicit association test (IAT). *European Review of Social Psychology, 17*, 74–147. https://doi.org/10.1080/10463280600681248

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of motor system in language comprehension. *The Quarterly Journal of Experimental Psychology, 61*, 825–850.

Fishbach, A., Shah, J. Y., & Kruglanski, A. W. (2004). Emotional transfer in goal systems. *Journal of Experimental Social Psychology, 40*, 723–738. https://doi.org/10.1016/j.jesp.2004.04.001

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*, 522–559. https://doi.org/10.1037/pspa0000160

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Sage http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

French, A. R., Franz, T. M., Phelan, L. L., & Blaine, B. E. (2013). Reducing Muslim/Arab stereotypes through evaluative conditioning. *The Journal of Social Psychology, 153*, 6–9. https://doi.org/10.1080/00224545.2012.706242

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation, 44*, 312–325. https://doi.org/10.1016/j.lmot.2013.03.003

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment, 24*, 218–225. https://doi.org/10.1027/1015-5759.24.4.218

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the implicit association test. *Journal of Consumer Research, 35*, 178–188. https://doi.org/10.1086/527341

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

Gollwitzer, P. M. (2015). Setting one's mind on action: Planning out goal striving in advance. In R. A. Scott, & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–14). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118900772.etrds0298.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69–119. https://doi.org/10.1016/S0065-2601(06)38002-1

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the implicit association test: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General, 134*, 420–425. https://doi.org/10.1037/0096-3445.134.3.420

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41. https://doi.org/10.1037/a0015575

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*, 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research, 1428*, 71–79. https://doi.org/10.1016/j.brainres.2011.06.026

Hofmann, W., Deutsch, R., Lancaster, K., & Banaji, M. R. (2010). Cooling the heat of temptation: Mental self-control and the automatic evaluation of tempting stimuli. *European Journal of Social Psychology, 40*, 17–25. https://doi.org/10.1002/ejsp.708

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369–1385. https://doi.org/10.1177/0146167205275613

Hollands, G. J., Prestwich, A., & Marteau, T. M. (2011). Using aversive images to enhance healthy food choices and implicit attitudes: An experimental test of evaluative conditioning. *Health Psychology, 30*, 195–203. https://doi.org/10.1037/a0022261

Kinoshita, S., & Peek-O'Leary, M. (2005). Does the compatibility effect in the race implicit association test reflect familiarity or affect? *Psychonomic Bulletin & Review, 12*, 442–452. https://doi.org/10.3758/BF03193786

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility–A model and taxonomy. *Psychological Review, 97*, 253–270. https://doi.org/10.1037/0033-295X.97.2.253

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*, 194–213. https://doi.org/10.1037/xge0000239

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S. Nosek, B. A., … (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M. Nosek, B. A., … (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001–1016. https://doi.org/10.1037/xge0000179

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments. (R package version 4.4-0)*. https://CRAN.R-project.org/package=ez.

Lebens, H., Roefs, A., Martijn, C., Houben, K., Nederkoorn, C., & Jansen, A. (2011). Making implicit measures of associations with snack foods more negative through evaluative conditioning. *Eating Behaviors, 12*, 249–253. https://doi.org/10.1016/j.eatbeh.2011.07.001

Levey, A. B., & Martin, I. (1975). Classical conditioning of human evaluative responses. *Behaviour Research and Therapy, 13*, 221–226.

Martiny-Huenger, T., Martiny, S. E., & Gollwitzer, P. M. (2015). Action control by if-then planning: Explicating the mechanisms of strategic automaticity in regard to objective and subjective agency. In B. Eitam, & P. Haggard (Eds.), *The sense of agency*. Oxford University Press.

Martiny-Huenger, T., Martiny, S. E., Parks-Stamm, E. J., Pfeiffer, E., & Gollwitzer, P. M. (2017). From conscious thought to automatic action: A simulation account of action planning. *Journal of Experimental Psychology: General, 146*, 1513–1525. https://doi.org/10.1037/xge0000344

Mattavelli, S., Van Dessel, P., & De Houwer, J. (2021). Why does the IAT reveal a preference for stimuli said to be paired with an unpleasant sound?Stalking the unexpected. *Collabra: Psychology, 7*, 18733. https://doi.org/10.1525/collabra.18733

Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the implicit association test. *Learning and Motivation, 34*, 203–217. https://doi.org/10.1016/S0023-9690(03)00003-1

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*, 413–417. https://doi.org/10.1111/1467-9280.00376

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford University Press.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139–165. https://doi.org/10.1037/0096-3445.133.2.139

Schmidt, J. R., De Houwer, J., & Rothermund, K. (2016). The parallel episodic processing (PEP) model 2.0: A single computational model of stimulus-response binding, contingency learning, power curves, and mixing costs. *Cognitive Psychology, 91*, 82–108. https://doi.org/10.1016/j.cogpsych.2016.10.004

Smith, C. T., Calanchini, J., Hughes, S., Van Dessel, P., & De Houwer, J. (2020). The impact of instruction- and experience-based evaluative learning on IAT performance: A quad model perspective. *Cognition and Emotion, 34*, 21–41. https://doi.org/10.1080/02699931.2019.1592118

Smith, E. R., & Conrey, F. R. (2007). Mental representations are states, not things: Implications for implicit and explicit measurement. In B. Wittenbrink, & N. Schwarz (Eds.), *Implicit measures of attitudes* (p. 247).

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34*, 1332–1345. https://doi.org/10.1177/0146167208321269

Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science, 20*, 192–196. https://doi.org/10.1177/0963721411408771

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology, 67*, 289–314. https://doi.org/10.1146/annurev-psych-122414-033417