Test Difficulty, True-Score Variance, and Predictive Power of Implicit Association Tests: Effects of Target Exemplar Stimuli

Merlin Urban, Tobias Koch, & Klaus Rothermund

FSU Jena, Germany

Collabra: Psychology, in press

Abstract

Following the test difficulty account, we manipulated the valence of the target exemplars of attitude Implicit Association Tests (IATs) with target categories that are generally considered positive or negative, using target exemplars that are either typically valenced (i.e., matching the respective target category's valence) or atypically valenced (i.e., deviating from the respective target category's valence). In doing so, we aimed to shift the IAT test difficulty from extreme test difficulty (i.e., mean IAT scores deviating strongly from zero) in the case of typically valenced exemplars to less extreme test difficulty (i.e., mean IAT scores deviating less strongly from zero) in the case of atypically valenced exemplars to increase the true-score variance and the predictive power of the IATs. We conducted three, pre-registered experiments (total n = 342) to test our hypotheses. We developed attitude IATs with the target categories environmental protection/environmental degradation, using both typically and atypically valenced target exemplars for each category. All data was analyzed using structural equation models. As hypothesized, in all three experiments the atypically compared to the typically valenced exemplars led to significantly less extreme test difficulties. However, contrary to our hypotheses, the less extreme test difficulty for IATs with atypical target exemplars did not result in a significant increase in true-score variance or predictive power in any of the experiments.

Keywords: Implicit Association Test (IAT), test difficulty, predictive power, target exemplar stimuli

The IAT has been criticized for its low predictive power, that is, its limited capacity to predict relevant outcome variables since years (e.g., Blanton et al., 2009; Meissner et al., 2019; Oswald et al., 2013), and the criticism is not subsiding (e.g., Corneille & Gawronski, 2024; Machery, 2022; Schimmack, 2021a, 2021b). Inspired by this criticism, however, attempts to address this problem continue to emerge. A recent and promising attempt is the test difficulty account (Urban et al., 2024). Urban et al. (2024) showed that, in line with classical test theory (CTT), IATs of moderate test difficulty tend to have more true-score variance and, consequently, more predictive power than IATs of extreme test difficulty. Thus, they argued that researchers intending to use the IAT to predict outcome variables should develop IATs of moderate test difficulty in order to increase their true-score variance.

The IAT is a computer-based task in which participants are asked to categorize exemplar stimuli displayed on the screen successively to one of four categories. The four categories comprise two target categories (e.g., environmental protection and environmental degradation) and two attribute categories (e.g., positive and negative). To categorize the exemplars, only two response keys are available, so that one target category and one attribute category always share a response key. There are two critical blocks that differ in their pairing of target and attribute categories, that share a response key (for a more detailed description of the overall block structure of IATs, see the Methods section). For example, in one block "environmental protection" and "positive" share one key, while "environmental degradation" and "negative" share the other key. In the other block, the target categories are reversed so that "environmental protection" and "negative" share one key, while "environmental degradation" and "positive" share the other key. The average response time difference between these two blocks produces the average IAT effect, which is typically interpreted as an indication of the strength of the associations between the target and attribute categories.

Within the test difficulty account (Urban et al., 2024), IAT test difficulty is defined based on the test difficulty concept of CTT: In this sense, IAT test difficulty reflects the extent to which people respond in the keyed direction of the theoretical construct. For example, in an attitude IAT, test difficulty indicates the extent to which people respond in favor of the attitude construct. The average IAT effect indicates IAT test difficulty. Given the structure of the IAT – its relative nature and block-based design – two conditions must be met in order to interpret the average IAT effect in terms of test difficulty: 1) The target category that functions as the attitude object in which keyed direction it is answered or not must be defined (referred to as the *relevant target category*), and 2) The block in which this target category is paired with the attribute category expressing the keyed direction must be defined (e.g., the positive attribute in a typical attitude IAT). Consider an environmental protection/environmental degradation attitude IAT. Say we define "environmental protection" as the relevant target category and the block in which "environmental protection" and "positive" share a response key as the subtrahend in the calculation of the IAT effect. Most likely, we would find a large positive average IAT effect, which would indicate the IAT to be an easy test. The test would be easy because participants would have responded strongly in the keyed direction of the theoretical construct (i.e., on average, they would have responded faster when the relevant target category "environmental protection" and the attribute "positive" shared a response key). According to the test difficulty account, this IAT should have very little true-score variance, as all participants would have responded in the same way and there would therefore be no interindividual differences, which ultimately reduces the predictive power of the IAT. Now, consider a Black/White attitude IAT. Say we define "Black" as the relevant target category and the block in which "Black" and "positive" share a response key as the subtrahend in the calculation of the IAT effect. Most likely, we would find a large negative average IAT effect, which would indicate the IAT to be a difficult test. The test would be difficult because participants would have responded strongly in opposition to the keyed direction of the theoretical construct (i.e., on average, they would have responded slower when the relevant target category "Black" and the attribute "positive" shared a response key). For the same reasons as for the environmental protection/environmental degradation IAT described above, the true-score variance and, consequently, the predictive power should be reduced. Finally, consider a Democrats/Republicans attitude IAT. Say we define “Democrats” as the relevant target category and the block in which “Democrats” and “positive” share a response key as the subtrahend in the calculation of the IAT effect. Most likely, we would find an average IAT effect close to or at zero, which would indicate the IAT to be a moderately difficult test. The IAT would be moderately difficult because participants would have responded neither more strongly in the keyed direction nor in the opposite direction (i.e., on average, they would have neither responded faster nor slower when the relevant target category “Democrats” and the attribute “positive” were paired on the same response key). This IAT should have a high amount of true-score variance, as some participants evaluate Democrats more positively than Republicans, while others evaluate Republicans more positively than Democrats and there would be considerable interindividual differences, which ultimately increases the predictive power of the IAT.[1]

Consequently, the test difficulty account argues against the common misconception that large IAT effects are desirable for correlational research (see for example Axt et al., 2021; Greenwald et al., 1998; Kurdi & Banaji, 2017, for such a rationale), because large IAT effects according to the test difficulty account represent IATs of extreme test difficulty, which are more likely to be associated with low true-score variance and low predictive power. Instead, the account argues in favor of using IATs with average IAT effects that are close to zero, that is, IATs of moderate test difficulty. However, this raises the question of how IATs of moderate test difficulty can be developed?

Urban et al. (2024) proposed three approaches for developing IATs with moderate test difficulty by modifying the IAT design: Manipulating the target reference category,

[1] Note that IAT test difficulty is sample-dependent and that we would assume the described results given a representative sample from the US.

manipulating the attribute categories, and manipulating the exemplar stimuli of the target categories. They presented evidence that the selection of a reference category with similar valence to the target category of interest shifts the IAT towards moderate test difficulty and that this shift in test difficulty is also associated with an increase in true-score variance and predictive power. In another study, Urban et al. (2025) showed that univalent (e.g., good vs. very good) compared to bivalent attribute categories (good vs. bad) also shift the IAT towards moderate test difficulty. Contrary to expectations, however, this shift in test difficulty had no effect on the true-score variance and the predictive power of the IAT, most likely because the use of univalent attributes compromises the construct validity of IATs.

In this study we will explore the third approach, which consists of manipulating the valence of the target exemplar stimuli of attitude IATs. For attitude IATs with target categories that can be regarded as generally positive or negative, we compare typically valenced with atypically valenced target exemplar stimuli. *Typically valenced* target exemplars match the general evaluations of the respective target category and thus have a clear positive (negative) valence for positive (negative) target categories (e.g., a positively evaluated exemplar such as “nature reserve” for a positively evaluated category such as “environmental protection”). *Atypically valenced* target exemplars deviate from the general evaluations of the respective target category in such a way that their valence is less extreme in the direction of the general valence of the respective target category and is therefore less positive, neutral, or negative for positive target categories and less negative, neutral, or positive for negative target categories (e.g., a less positively evaluated exemplar such as “climate tax” for a positively evaluated category such as “environmental protection”).

Previous research on the influence of target exemplar valence

Several studies have shown that typically valenced target exemplars lead to larger absolute average IAT effects (i.e., IATs of more extreme test difficulty), whereas atypically valenced target exemplars produce absolute average IAT effects that are smaller and closer to zero, that is, resulting in IATs with more moderate test difficulty (e.g., Bluemke & Friese, 2006; Gast & Rothermund, 2010; Govan & Williams, 2004). Bluemke and Friese (2006) conducted two experiments with West German participants and found large absolute average IAT effects for East vs. West IATs when typically valenced target exemplars were used, and smaller absolute average IAT effects when atypically valenced target exemplars were used. Govan and Williams (2004) found a large absolute average IAT effect for a Flower vs. Insect IAT when typically valenced target exemplars were used, compared to an IAT with atypically valenced target exemplars. Gast and Rothermund (2010) found a similar pattern of results for an Old vs. Young IAT. It should be noted that some studies found no differences between typical and atypical target exemplars (De Houwer, 2001), but follow up research provided convincing explanations for these null results (see Bluemke & Friese, 2006).

Consequently, our hypothesis regarding the influence of target exemplar valence on IAT test difficulty, that is, on the average IAT effect, is not new, but has already been demonstrated. What is new, however, are our hypotheses that IAT test difficulty should also have downstream effects on the true-score variance and the predictive power of the IAT. While existing literature has investigated the influence of target exemplar valence on the average IAT effect and on possible underlying processes that mediate the effect (e.g., cross category associations, Bluemke & Friese, 2006; category redefinition, Govan & Williams, 2004), none of the previous studies investigated the influence of target exemplar valence on the true-score variance and the predictive power of IATs. In fact, as far as we know, no study has yet investigated the influence of target exemplar valence on any specific psychometric property of the IAT. There are two studies that have examined the effect of attribute exemplar stimuli, but not of target exemplar valence, on specific psychometric properties of the IAT. Stieger et al. (2010) showed that individualizing the attribute exemplar selection by asking participants to self-select the attribute exemplars compared to using the standard procedure of attribute exemplar selection, in which the attribute exemplars are selected by the researchers, did not influence the internal consistencies, retest-reliabilities, implicit-implicit (I-I) and implicit-explicit (I-E) correlations of the IATs. Axt et al. (2021) showed that the variability of attribute exemplars typically used in attitude IATs had no influence on the average IAT effect (i.e., IAT test difficulty), internal consistencies, I-E correlations and prediction of known-group differences.[2]

As no study to date has investigated the influence of target exemplar valence on specific psychometric properties of the IAT, we test the relations between IAT test difficulty, true-score variance, and predictive power derived from the test difficulty account, thereby evaluating the potential utility of manipulating target exemplar valence for optimizing the IAT as a diagnostic tool. Furthermore, we also make an initial contribution to understanding whether and how target exemplar valence influences specific psychometric properties of the IAT.

Hypotheses and overview of the experiments

Our previous considerations can be summarized in the following hypotheses: Using atypically rather than typically valenced target exemplar stimuli in the case of an IAT with target categories that are a priori thought to be unambiguously positive or negative a) shifts the test difficulty from a more extreme to a less extreme test difficulty, that is, shifts the IAT effect from being strongly different from zero to being less strongly different from zero (H1), b) increases the true-score variance (H2), and c) increases the predictive power of the resulting IAT (H3).

We tested our hypotheses using different experimental designs and IAT procedures while we always used the same attitude domain environmental protection/environmental degradation. For all experiments we compared typically and atypically valenced target

[2] Hogenboom et al. (2024) also tested the validity of both target and attribute exemplar stimuli. They did not, however, investigate the influence of the exemplars on the psychometric properties of the IAT, but instead only investigated whether exemplars elicited fast (< 800 ms) and accurate (< 10% errors) responses (as suggested by Greenwald et al., 2022).

exemplars, whereby we selected the valence of the atypical exemplars in such a way that they differed as much as possible from the valence of the respective target categories, since we expected this to have the strongest effect on IAT test difficulty, true-score variance and predictive power.[3]

In Experiment 1, we manipulated the valence of the target exemplars within a single IAT so that the IAT contained a typically and an atypically valenced target exemplar stimulus set. As was already discussed by Bluemke and Friese (2006), one main advantage of a within design in this context is that the probability of so-called subtyping processes due to the manipulation of the valence of the exemplar stimuli (e.g., a redefinition of the target categories; see Govan & Williams, 2004), which could call into question the measurement of the intended constructs, is reduced to a minimum. This is because the entirety of all exemplar stimuli for the target categories are presented randomly within each test block, making it difficult to mentally activate specific subtypes of the target categories, as this would require a more coherent presentation of the respective exemplars.

In Experiment 2, to test the robustness of the results of Experiment 1, we again used a within design, but made minor changes to the procedural design of the IAT. Specifically, we decreased the number of exemplars per category and increased the number of trials per exemplar in the hope of increasing the overall reliability, and we adjusted the selected target exemplars.

In Experiment 3, we switched from a within to a between design, resulting in a typical and an atypical IAT with only typical and only atypical valenced target exemplars, respectively. While the within design reduces the likelihood of subtyping processes such as a redefinition of the categories, there is a risk that the resulting IAT might become more challenging for the participants, since the valence of the exemplar stimuli varies not only between but also within

[3] Note, however, that in none of the experiments the valence of the atypical exemplars was opposite to the valence of the respective target category, since the few exemplars that were rated as such were no longer rated as representative of the respective category.

the two target categories. As a result, responding may become less spontaneous and more controlled, possibly reducing the reliability and counteracting effects of the manipulation. These potential risks should be eliminated in a between design in which the valence of the exemplars within a target category is more similar. Since both types of experimental design have different advantages and disadvantages, the use of both types seems to allow for a more comprehensive and rigorous evaluation of our hypotheses.

To examine the predictive power of the IATs, we used both direct attitude measures and behavioral measures as outcome variables for all three experiments, i.e., we compared I-E and implicit-criterion (I-C) correlations between the atypical and the typical stimulus set in Experiments 1 and 2, and between the atypical and the typical IAT in Experiment 3.

Transparency and Openness

The research was granted ethical approval by the University of xxx Ethics Committee. All materials related to the following studies, including preregistrations, stimulus materials, raw data, curated data, and code, are available publicly on our OSF project page (Link: https://osf.io/3meza/?view_only=dcdfd55cc56a45d6929c3d391dd0ce2c). We disclose all measures, manipulations, data exclusions, and how we determined the sample size of the following studies. All analyses were performed using R (version 4.2.2, R Core Team, 2021). We used the R package lavaan (Rosseel, 2012) for our main analyses.

Experiment 1

In Experiment 1 we developed an attitude IAT with the target categories "environmental protection" and "environmental degradation" expecting the former category to generally elicit positive and the latter category to generally elicit negative evaluations. We then manipulated the valence of the target stimuli within the IAT to have a typical valence (typical stimulus set) or an atypical valence (atypical stimulus set).

Methods

Design and Procedure

We used a mixed design with the within factor *target exemplar valence*, which had two levels (typically valenced vs. atypically valenced stimulus set) and the between factor *IAT block order*, which also had two levels (compatible vs. incompatible block first). Participants were randomly assigned to one of the block orders. They were then asked to provide demographic data and to rate the representativeness of the target exemplar stimuli for the target categories. Subsequently, the IAT was administered and, in a final step, participants were asked to complete a questionnaire measuring their environmental attitudes and behaviors.

Sample

The total sample consisted of 155 participants and was collected via mailing lists of the University of xxx and via social media channels. Students of the University of xxx received course credits. Three participants were excluded – one for not being a native German speaker and two for data quality reasons (see the results section for a detailed description of the exclusion criteria). As a result, we were left with a final sample of 152 participants (84% female; 99% in educational training; 96% psychology students; mean age of *M* = 21.9 years [*SD* = 2.84]), exceeding the targeted 100 participants, based on a one-tailed a priori power analysis for z-tests of two dependent correlations with a common index with G*Power (alpha = .05, power = .8, rho1 = .0, rho2 = .3, rho3 = .3).

Measures

IAT. The target categories of the IAT were “environmental protection”/“environmental degradation”. Per target category we used five typically and five atypically valenced exemplar stimuli. The typical exemplars for environmental protection were positive (i.e., biodiversity, nature reserve, saving energy, planting trees, sustainability) and for environmental degradation negative words (i.e., water pollution, global warming, deforestation, climate crisis, waste pollution) while the atypical exemplars for environmental protection were less positive (i.e., climate tax, radical environmental activists, climate strike, ecos, consumption abstinence) and for environmental degradation less negative words (i.e., economic growth, driving, meat-based diet, globalization, world travel). We selected the exemplars based on a pretest with 102 participants in which we asked them to rate the valence of 157 environmentally related word exemplars on a 9-point bipolar scale with endpoints ranging from 1 (*extremely negative*) to 9 (*extremely positive*) and the representativeness of the exact same exemplars for the target categories on a 6-point bipolar scale with endpoints ranging from 1 (*strong connection with environmental degradation*) to 6 (*strong connection with environmental protection*). Alongside using the representativeness and the valence ratings of the pretest as target exemplar selection criteria (see Supplement 1 for a more detailed description of the selection criteria and descriptive statistics), we applied the recommended criteria (Greenwald et al., 2022). The selection of the attribute exemplars, on the other hand, was based only on the recommended criteria. The exemplars for the attribute category "good" were ten positive adjectives (e.g., good, pleasant, happy) and for the attribute category "bad" ten negative adjectives (e.g., flawed, bad, awful).

With regard to the IAT block structure and the number of trials per block, we followed the standard IAT protocol described in Greenwald et al. (2022), which resulted in 20, 20, 40, 80, 40, 40, and 80 trials for the respective seven blocks. Consequently, each exemplar stimulus within a block containing the respective category was presented once or twice, depending on the block. Target and attribute exemplar stimuli were randomly presented in black font on a white background in each block. Participants were asked to respond as quickly and accurately as possible by pressing the left (*D*) or the right response key (*L*) on their computer keyboard. If they took longer than three seconds to respond or made a mistake, they received a corresponding feedback message in red font with the instruction to proceed by pressing the correct response key.

IAT effects were calculated based on the D score algorithm (Greenwald et al., 2003). Only the target stimuli were included in the calculation, as we were interested in comparing the pure effect of the atypical vs. the typical stimulus set (see Gast & Rothermund, 2010). D scores were coded in such a way that positive D scores indicate a more positive evaluation of the target category environmental protection.

Outcome variables. We collected two direct attitude measures and one behavior measure. The first direct attitude measure consisted of questionnaire items that assessed the evaluation of all target exemplar stimuli (hereafter referred to as exemplar evaluation measure). Each exemplar was to be rated on a 9-point bipolar scale with endpoints ranging from 1 (*extremely negative*) to 9 (*extremely positive*). We recoded the exemplars of the target category environmental degradation in such a way that higher scores on the measure indicate a more positive/negative evaluation of environmental protection/environmental degradation (internal consistency of $\omega_t = .88$). The second direct attitude measure consisted of questionnaire items that assessed gut reactions and actual feelings towards the target categories (hereafter referred to as target evaluation measure). The items were: (a) “Rate your gut reactions towards environmental protection”, (b) “Rate your actual feelings towards environmental protection”, (c) “Rate your gut reactions towards environmental degradation”, and (d) “Rate your actual feelings towards environmental degradation”. All items were to be rated on a 10-point bipolar scale with endpoints ranging from 1 (*extremely negative*) to 10 (*extremely positive*). We calculated difference scores between the ratings of the two target categories, once for gut reactions and once for actual feelings, with higher scores indicating a more positive/negative evaluation of environmental protection/environmental degradation (internal consistency of $\omega_t = .70$). The behavior measure consisted of two questionnaire items assessing environmental behavior: (a) “I behave environmentally friendly” and (b) “I behave environmentally harmful”. Both items were to be rated on a 7-point frequency scale with endpoints ranging from 1 (*never*) to 7 (*very often*). We calculated difference scores in such a way that higher values indicate more/less environmentally friendly/environmentally harmful behavior (internal consistency of $\omega_t = .44$).[4]

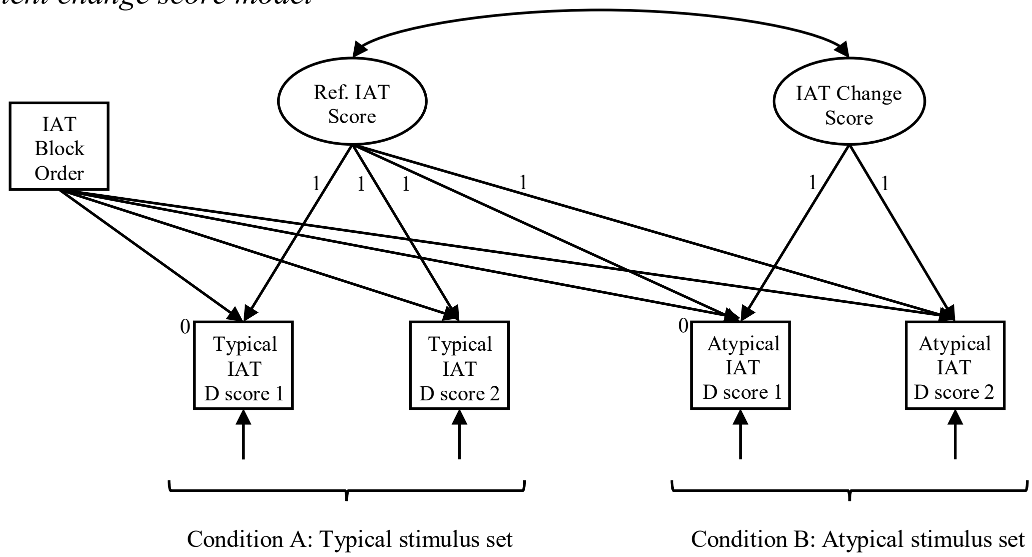
Data analysis

We ran structural equation models (SEM) to test our hypotheses in a single, unified statistical framework. To examine whether the typical and atypical stimulus set differed as expected in terms of their latent means to determine their test difficulties (H1) and their latent variances to determine their true-score variances (H2), we used latent change score modeling to account for the repeated measurement design. For a conceptual representation of the model see Figure 1. The model consisted of two correlated latent variables: The reference (or baseline) latent D score variable, representing the IAT scores for the typical stimulus set, and the latent D score change variable, representing the difference in IAT scores between the typical and atypical stimulus sets. The manifest indicators for the different conditions were two D scores for each condition, calculated from the short test block and the long test block of the IAT. To control for potential block order effects, we regressed the manifest D score indicators on the manifest covariate IAT block order, assuming an equal effect of IAT block order on the two conditions. The manifest covariate was centered at the grand mean before the analysis. When applying this latent change score model, the variance of the latent D score change variable was negative and not significantly different from zero. This result has practical implications for H2 which we describe in more detail below. In the further application, we fixed the variance of the latent D score change variable and the covariance between the latent D score change variable and the reference latent D score variable to zero. This restricted model allowed us to test overall mean differences (i.e., difference in test difficulties) between both experimental conditions, but assumed no interindividual differences.

[4] Note that the internal consistency refers to the two items and not to the difference variable.

Figure 1

Latent change score model

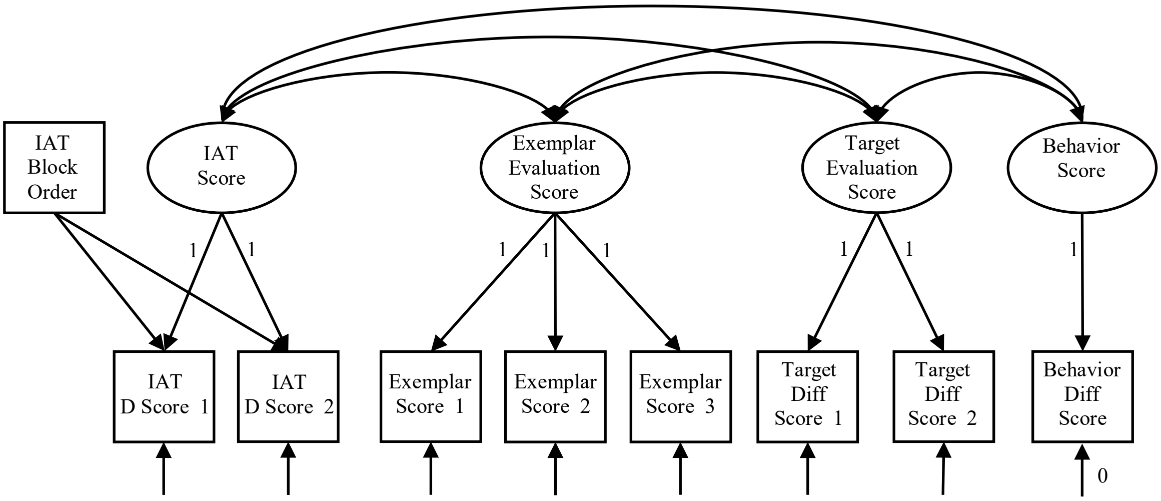


Note. Circles represent latent and rectangles observed variables. IAT = implicit association test; Ref. = reference.

To test whether the typical and atypical stimulus set differed as expected in terms of their latent correlations to determine their predictive power (H3) we used latent state modeling. We tested the predictive power of the two stimulus sets in separate models. Thereby, we circumvented multicollinearity issues caused by a high correlation between the two stimulus sets within a single model, a result that has practical implications for H3 and that we describe in more detail below. For a conceptual representation of the model that was fitted for both stimulus sets see Figure 2. The model consisted of four correlated latent variables: The latent D score variable measured via two manifest indicators, i.e., D scores calculated based on the short and the long test block; the latent exemplar evaluation variable measured via three manifest indicators, i.e., three parcels that were created by aggregating the evaluations of the target exemplar stimuli; the latent target evaluation variable measured via two manifest indicators, i.e., difference scores once based on the gut reactions and once based on the actual feelings; and the latent behavior variable measured via one manifest indicator, i.e., difference scores based on the two questionnaire items assessing the behavior. We again controlled for potential block order effects in a similar manner as already described above with the only difference that this time no equal effect on the two conditions could be assumed because we ran two separate models.

Figure 2

Latent state model that was fitted for both the typical and the atypical stimulus set



Note. Circles represent latent and rectangles observed variables. IAT = implicit association test; Target Diff Score 1 = difference score of the target evaluation measure based on gut reactions; Target Diff Score 2 = difference score of the target evaluation measure based on actual feelings; Behavior Diff Score = behavior difference score.

Results

Preliminary analyses

Ensuring data quality. As we used the D score algorithm to compute IAT effects, we excluded participants who responded faster than 300 ms in 10% or more of the trials across all test blocks (1.3% of participants), and we excluded responses with latencies exceeding 10,000 ms (0.04% of the trials).

Descriptive statistics, multivariate normal distribution, handling missing values, and measurement invariance. Descriptive statistics for all manifest indicators are provided in Supplement 1 on our OSF project page.[5] Neither the manifest indicators for the latent

[5] We do not include descriptive statistics such as mean values, standard deviations and correlations in the main text because we report precisely these parameters in our SEM analyses.

change score model were multivariate normally distributed (Mardia's skewness = 59.40, $p <$ .001; Mardia's kurtosis = 1.51, $p = .13$), nor the manifest indicators for the latent state models were multivariate normally distributed (Mardia's skewness$_{\text{atypical}}$ = 258.38 $p < .001$; Mardia's kurtosis$_{\text{atypical}}$ = 5.36, $p < .001$; Mardia's skewness$_{\text{typical}}$ = 281.47, $p < .001$; Mardia's kurtosis$_{\text{typical}}$ = 5.51, $p < .001$), which is why we used the maximum likelihood mean-variance adjusted (MLMV) estimator for all models. None of the manifest indicators had missing values. We assumed strict measurement invariance (MI) for the change score model as the strict MI model had an excellent fit, S-B $\chi^2_{\text{strict MI}}$ (10) = 11.53, $p = .32$; RMSEA$_{\text{strict MI}}$ = 0.03; CFI$_{\text{strict MI}}$ = 0.99; SRMR$_{\text{strict MI}}$ = 0.05; AIC$_{\text{strict MI}}$ = 488.23; BIC$_{\text{strict MI}}$ = 512.42 (see Supplement 1 for a more detailed explanation of the MI analysis). We could not test the latent state models for MI because we ran a separate model for each stimulus set, i.e., for each of the experimental conditions, but the overall model fits of the atypical and typical latent state models were very good, S-B χ^2_{atypical} (22) = 28.48, $p = .16$; RMSEA$_{\text{atypical}}$ = 0.05; CFI$_{\text{atypical}}$ = 0.98; SRMR$_{\text{atypical}}$ = 0.06, and S-B χ^2_{typical} (22) = 24.57, $p = .32$; RMSEA$_{\text{typical}}$ = 0.03; CFI$_{\text{typical}}$ = 0.99; SRMR$_{\text{typical}}$ = 0.06, respectively.

Main analyses

H1: Test difficulty. As hypothesized the latent mean of the atypical stimulus set was not only descriptively smaller, $\hat{\mu}_{\text{atypical}} = 0.718$, compared to the latent mean of the typical stimulus set, $\hat{\mu}_{\text{typical}} = 0.809$, but this difference was also statistically significant, $z = -3.79$, $p <$ .001 (see Table 1 for the latent means and their standard errors). However, the IAT effect of the atypical stimulus set was still large. In fact, the latent mean value of the atypical stimulus set was significantly different from zero, $z = 23.58$, $p < .001$.

Table 1

Latent Means, Latent True-Score Variances, and Reliabilities for the Typical and Atypical Stimulus Set in the Strict Invariance Change Score Model as well as Latent Correlations for the Typical and Atypical Stimulus Set in the latent state models (Experiment 1)

Condition	L mean (*SE*)	L variance (*SE*)	L cor target measure (CI)	L cor exemplar measure (CI)	L cor behavior measure (CI)	Reliability
Typical Stimulus Set	0.81 (.03)	0.038 (.008)	-.29 (-.77, .16)	-.12 (-.53, .28)	-.28 (-.69, .03)	.39
Atypical Stimulus Set	0.72 (.03)	0.038 (.008)	.08 (-.26, .42)	.16 (-.13, .48)	-.33 (-.59, -.06)	.39

Note. L = latent; cor = correlation; CI = bootstrap-bias-corrected confidence intervals.

H2: True-score variance. Contrary to our hypothesis, the true-score variances of the atypical and typical stimulus set did not differ but was identical in size. This was a direct mathematical consequence of the fact that the variance of the latent D score change variable was empirically zero and could also be fixed to zero without worsening the model fit. The model fit indices were as follows: $\text{AIC}_{\text{free variance}} = 489.92$, $\text{BIC}_{\text{free variance}} = 520.16$, compared to $\text{AIC}_{\text{fixed variance}} = 488.23$ and $\text{BIC}_{\text{fixed variance}} = 512.42$. The zero variance of the difference score variable means that there are no interindividual differences in the latent change from the typical to the atypical stimulus set and that the sets only differ by a constant (i.e., equal) shift of the IAT scores for all participants. Because of this perfect linear dependency between the typical and atypical IAT scores, the variance of the IAT scores has to be identical in both experimental conditions, $\hat{\sigma}_{\text{atypical}}{}^2 = \hat{\sigma}_{\text{typical}}{}^2 = 0.038$ as the variance does not change when a constant is added (see Table 1 for the true-score variances and their standard errors).

H3: Predictive power. Contrary to our hypothesis the latent correlations of the atypical stimulus set were not significantly larger than those of the typical stimulus set. This follows from the fact that the stimulus sets were perfectly correlated with each other when modeled in a single model. In other words, the stimulus sets cannot be statistically separated from each other and formed a single common factor that makes the same predictions. To be able to report correlations for the atypical and typical stimulus set independently, we modeled the stimulus sets in separate models. The latent correlations of the atypical and typical stimulus set with the target evaluation measure were, $\hat{r}_{\text{atypical}} = .08$, $z = 0.52$, $p = .6$ and $\hat{r}_{\text{typical}} = -.29$, $z = -1.39$, $p = .16$, respectively, with the exemplar evaluation measure they were $\hat{r}_{\text{atypical}} = .16$, $z = 1.19$, $p = .23$ and $\hat{r}_{\text{typical}} = -.12$, $z = -0.70$, $p = .48$, respectively, and with the behavior measure they were $\hat{r}_{\text{atypical}} = -.33$, $z = -2.66$, $p < .01$ and $\hat{r}_{\text{typical}} = -.28$, $z = -1.84$, $p = .07$, respectively (see Table 1 for the latent correlations and *CIs*). It should be noted that the above correlations are provided for completeness only. Since both stimulus sets are perfectly correlated and form a common factor when included in the same model, the above correlations do not differ statistically, and any numerical differences are due to sample variation.[6]

Discussion

In Experiment 1, we manipulated the valence of the target exemplar stimuli of an IAT with the target categories environmental protection/environmental degradation, creating a typical and an atypical stimulus set for the two target categories in order to influence the test difficulty, true-score variance, and predictive power of the IAT. In line with our hypotheses using atypical exemplars instead of typical exemplars (a) resulted in a shift from a more extreme to a less extreme IAT test difficulty, resulting in a smaller average IAT effect which, however, was still not close to zero and thus not of moderate difficulty. Contrary to our hypotheses, this did not (b) lead to an increase in true-score variance of the IAT nor (c) to an increase in its predictive power as assessed by the I-E or I-C correlations. In Experiment 2, we further modified the IAT procedure in the hope of increasing the somewhat low reliabilities of the effects for the two stimulus sets, and to get a second, independent test of our hypotheses.

Experiment 2

In Experiment 2 we again developed an Environmental protection/Environmental degradation attitude IAT with an atypical and a typical target exemplar stimulus set, but made

[6] Note that the negative correlation with the behavioral measure is surprising, but that this effect does not appear to be particularly stable, as it disappears if one does not control for block order, and as it was not found again in the following experiments.

the following modifications to the material: a) We changed the IAT procedure by decreasing the number of exemplar stimuli per category, increasing the number of test blocks and, consequently, increasing the number of trials per exemplar stimulus in the hope to increase the overall reliability of the effects for the typical and atypical stimulus set, b) we adjusted the selection of the target exemplar stimuli, and c) we adapted the direct attitude measures to match the revised IAT (see the Measures section for more details).

Methods

Design and Procedure

The design and procedure of Experiment 2 was identical to that of Experiment 1 (see the corresponding section of Experiment 1 for a detailed description).

Sample

The total sample consisted of 97 participants, recruited and compensated in the same way as in Experiment 1. Two participants were excluded – one for not being a native German speaker and one for data quality reasons (see the Results section of Experiment 1 for exclusion criteria), resulting in a final sample of 95 participants (84% female; 98% in educational training; 97% psychology students; mean age of *M* = 21.7 years [*SD* = 2.63]). Due to resource limitations, the sample was smaller than the targeted 136 participants, based on a one-tailed a-priori power analysis for z-tests of two dependent correlations with a common index with G*Power (alpha = .05, power = .8, rho1 = .2, rho2 = .0, rho3 = .56).

Measures

IAT. As in Experiment 1, the target categories of the IAT were "environmental protection"/"environmental degradation" and the target exemplar stimuli for both categories were either typically or atypically valenced. In contrast to Experiment 1, we reduced and slightly adjusted the selected target exemplars: the typically valenced exemplars for environmental protection were three positive (i.e., nature reserve, solar energy, fair trade) and for environmental degradation three negative word exemplars (i.e., deforestation, waste pollution, microplastics), while the atypically valenced exemplars for environmental protection were three less positive (i.e., radical environmental activists, consumption abstinence, airplane ban) and for environmental degradation three less negative word exemplars (i.e., economic growth, luxurious lifestyle, air conditioning). We selected the exemplars based on the already described pretest and another pretest with 29 participants, in which we asked participants to rate the valence and representativeness of 29 additional environment-related word exemplars in the same way as in the first pretest. The selection criteria for the target exemplars remained the same as in Experiment 1 (see Supplement 2 for the specific descriptive statistics of the selected exemplars). As we reduced the number of target exemplars, we also reduced the number of attribute exemplars. We selected 6 positive exemplars for the attribute category "good" and 6 negative exemplars for the attribute category "bad" from the attribute exemplars already used in Experiment 1. Furthermore, we modified the IAT block structure and the number of trials per block as compared to Experiment 1 so that the IAT consisted of the following 13 blocks: the target discrimination practice block (12 trials), the attribute discrimination practice block (12 trials), initial combined test block (24 trials), reversed target discrimination practice block (24 trials), reversed combined test block (24 trials), followed by four repetitions of the two test blocks in alternating order in order to have multiple manifest indicators for the latent constructs, and to reduce the influence of block sequence effects (see Meissner & Rothermund, 2013). Consequently, each exemplar stimulus within a block was presented once. Apart from these changes, the IAT procedure remained the same as in Experiment 1.

IAT effects were calculated and coded in the same way as in Experiment 1, so that positive D scores indicate a more positive evaluation of the target category environmental protection.

Outcome variables. We used the same outcome variables as in Experiment 1 (exemplar evaluation measure, target evaluation measure, and behavior measure with internal consistencies of $\omega_t = .86$, $\omega_t = .51$, and $\omega_t = .40$, respectively), the only difference being that the target exemplars of the exemplar evaluation measure corresponded to the selected target exemplars of the IAT in Experiment 2 (see the corresponding section of Experiment 1 for more details).

Data analysis

Since the experimental design and hypotheses were identical to those of Experiment 1, we again used latent change score modeling to test H1 and H2, and latent state modeling to test H3. The main difference with regard to the statistical analyses compared to Experiment 1 was that we, in line with the modification of the IAT procedure, increased the number of manifest indicators for the IAT. Accordingly, the latent change score model and the latent state models of Experiment 2 differed from the respective models in Experiment 1 in that this time the latent D score variables were measured via five instead of two manifest indicators, i.e. five D-scores calculated on the basis of the five test blocks of the IAT.

As in Experiment 1, when applying the latent change score model, the variance of the latent D score change variable was negative and not significantly different from zero, so we again fixed the variance of the latent D-score change variable and the covariance between the latent D-score change variable and the reference latent D-score variable to zero in the further application. Also as in Experiment 1, we tested the predictive power of the two stimulus sets in separate models, as the two stimulus sets were perfectly correlated when being included in a single model.

Results

Preliminary analyses

Ensuring data quality. We applied the same criteria for ensuring data quality as in Experiment 1 based on which 1.04% of the participants and 0.05% of the trials were excluded (refer to the corresponding section of Experiment 1 for a detailed description).

Descriptive statistics, multivariate normal distribution, handling missing values, and measurement invariance. Descriptive statistics for all manifest indicators can be found in Supplement 2 on our OSF project page. The manifest indicators for the latent change score model violated the assumption of a multivariate normal distribution (Mardia's skewness = 291.73, $p < .001$; Mardia's kurtosis = 1.13, $p = .26$), which was also true for the manifest indicators for the latent state models (Mardia's $\text{skewness}_{\text{atypical}}$ = 334.57 $p = .03$; Mardia's $\text{kurtosis}_{\text{atypical}}$ = 1.81, $p = .07$; Mardia's $\text{skewness}_{\text{typical}}$ = 369.56, $p < .001$; Mardia's $\text{kurtosis}_{\text{typical}}$ = 2.53, $p = .01$). Accordingly, we used the MLMV estimator for all models. There were no missing values for any of the manifest indicators. We assumed strict MI for the change score model as the strict MI model had an excellent fit, S-B $\chi^2_{\text{strict MI}}$ (58) = 58.03, $p = .47$; $\text{RMSEA}_{\text{strict MI}}$ = 0.00; $\text{CFI}_{\text{strict MI}}$ = 1.00; $\text{SRMR}_{\text{strict MI}}$ = 0.09; $\text{AIC}_{\text{strict MI}}$ = 1699.4; $\text{BIC}_{\text{strict MI}}$ = 1742.8. The overall model fit of the atypical latent state model, S-B χ^2_{atypical} (49) = 53.94, $p = .29$; $\text{RMSEA}_{\text{atypical}}$ = 0.04; $\text{CFI}_{\text{atypical}}$ = 0.97; $\text{SRMR}_{\text{atypical}}$ = 0.06, and the typical latent state model, S-B χ^2_{typical} (49) = 45.52, $p = .62$; $\text{RMSEA}_{\text{typical}}$ = 0.00; $\text{CFI}_{\text{typical}}$ = 1.00; $\text{SRMR}_{\text{typical}}$ = 0.06, was excellent.

Main analyses

H1: Test difficulty. In line with our hypothesis, the latent mean of the atypical stimulus set was descriptively smaller, $\hat{\mu}_{\text{atypical}}$ = 0.659, than the latent mean of the typical stimulus set, $\hat{\mu}_{\text{typical}}$ = 0.881, and this difference was statistically significant, $z = -6.845$, $p < .001$ (see Table 2 for the latent means and their standard errors). Furthermore, the IAT effect of the atypical stimulus set was still significantly different from zero, $z = 12.943$, $p < .001$.

Table 2

Latent Means, Latent True-Score Variances, and Reliabilities for the Typical and Atypical Stimulus Set in the Strict Invariance Change Score Model as well as Latent Correlations for the Typical and Atypical Stimulus Set in the latent state models (Experiment 2)

Condition	L mean (*SE*)	L variance (*SE*)	L cor target measure (CI)	L cor exemplar measure (CI)	L cor behavior measure (CI)	Reliability
Typical Stimulus Set	0.88 (.05)	0.057 (.013)	-.17 (-.79, .25)	.07 (-.39, .45)	.07 (-.34, .45)	.48
Atypical Stimulus Set	0.66 (.05)	0.057 (.013)	.23 (-.24, .73)	.04 (-.32, .38)	-.05 (-.39, .27)	.48

Note. L = latent; cor = correlation; CI = bootstrap-bias-corrected confidence intervals.

H2: True-score variance. Contrary to our hypothesis, the true-score variance of the atypical stimulus set was identical to that of the typical stimulus set. As in Experiment 1, the variance of the latent D score change variable was empirically zero and did not lead to a worse model fit when fixed to be zero: $AIC_{\text{free variance}} = 1701.6$, $BIC_{\text{free variance}} = 1750.1$, compared to $AIC_{\text{fixed variance}} = 1699.4$, $BIC_{\text{fixed variance}} = 1742.8$. Again, this implies that the true-score variance of the atypical and typical stimulus set must be identical, $\widehat{\sigma}_{\text{atypical}}{}^{2} = \widehat{\sigma}_{\text{typical}}{}^{2} = 0.057$ (for a more detailed explanation of why the results imply equal true-score variances, see the corresponding section of Experiment 1; see Table 2 for the true-score variances and their standard errors).

H3: Predictive power. Contrary to our hypothesis, the latent correlations of the atypical stimulus set were not significantly larger than those of the typical stimulus set. As in Experiment 1, the stimulus sets were perfectly correlated when entered in one model. To be able to report the correlations for the two sets, we modeled the stimulus sets in separate models. The latent correlations of the atypical and typical stimulus set with the target evaluation measure were $\hat{r}_{\text{atypical}} = .23$, $z = 1.20$, $p = .23$, and $\hat{r}_{\text{typical}} = -.17$, $z = -0.80$, $p = .42$, respectively, with the exemplar evaluation measure they were $\hat{r}_{\text{atypical}} = .04$, $z = 0.23$, $p = .82$, and $\hat{r}_{\text{typical}} = .07$, $z = 0.37$, $p = .71$, respectively, and with the behavior measure they were $\hat{r}_{\text{atypical}} = -.05$, $z = -0.31$, $p = .76$, and $\hat{r}_{\text{typical}} = .07$, $z = 0.40$, $p = .69$, respectively (see Table 2 for the latent correlations and *CI*s; see Experiment 1 for an explanation of why there are descriptive differences even though the stimulus sets must make the same predictions).

Discussion

In Experiment 2, we again manipulated the valence of the target exemplar stimuli of an Environmental protection/Environmental degradation attitude IAT, creating a typical and an atypical stimulus set. As hypothesized the typical compared to the atypical stimulus set (a) had a more extreme IAT test difficulty, i.e., had a larger average IAT effect, but contrary to our hypotheses, the typical compared to the atypical stimulus set (b) did not have lower true-score variance nor (c) lower predictive power as assessed by the I-E or I-C correlations. Accordingly, although we used somewhat different target exemplars and modified the procedural design of the IAT, which led to an increase in the reliability of the IAT effects for the two stimulus sets, we found exactly the same pattern of results as in Experiment 1 with respect to our hypotheses. Consequently, the results appear to be robust. So far, however, we have only investigated our hypotheses by using a within design. Due to the mixed valence of the exemplar stimuli within the two target categories and the randomized presentation of the exemplars, in combination with the overall rather similar valence of both the typical and atypical exemplars, participants might not have been able to clearly distinguish between the typical and atypical exemplars on the basis of their valence. Due to this conflating of exemplars into a homogeneous overall impression, responses to both sets of stimuli might have been similar for the two target categories, resulting in more strategic responding, IAT effects of similar magnitude, and identical variances and correlations with criterion variables. This would not only explain the generally rather low reliabilities of the stimulus sets, it would also weaken the overall effect of the manipulation. It is therefore possible that the effect of target exemplar valence on test difficulty would be stronger in a between design (see also Bluemke & Friese, 2006, for an indication of such a pattern of results), which might also be better suited to investigate the hypothesized effects of our exemplar manipulations on true-score variance and predictive power. For these reasons, we switched to a between design in Experiment 3.

Experiment 3

In Experiment 3, we developed two Environmental protection/Environmental degradation attitude IATs, one with only typically valenced target exemplar stimuli (typical IAT) and one with only atypically valenced target exemplar stimuli (atypical IAT). In addition to changing the design of the experiment we also made the following changes to the material of Experiment 3 in comparison to Experiments 1 and 2: a) we modified the number of exemplar stimuli per category, b) we adjusted the selection of the target exemplar stimuli, and c) we adapted the direct attitude measures to match the respective IATs (see the Measures section of Experiment 3 for more details).

Methods

Design and Procedure

We used a between design with two factors: *target exemplar valence* as a two-level between factor (typical IAT, i.e., IAT with typically valenced exemplars vs. atypical IAT, i.e., IAT with atypically valenced exemplars) and *block order* as a two-level between factor (compatible vs. incompatible block first). Participants were randomly assigned to one of the conditions. They were then asked to provide demographic data, rate the representativeness of the target exemplar stimuli for the target categories, and complete the respective IAT. Afterwards, a category IAT (cf. Govan & Williams, 2004) was administered so that we could explore potential subtyping processes (e.g., a redefinition of the target categories). Finally, participants were asked to complete a questionnaire measuring their environmental attitudes and behaviors.

Sample

The total sample size comprised 100 participants. We recruited and compensated the participants in the same way as in Experiment 1 and 2. Three participants were excluded for not meeting the requirement of having German as a mother tongue, one participant was excluded for data quality reasons (see the Results section of Experiment 1 for a detailed explanation of the exclusion criteria), and finally, another participant was excluded who had taken part in the experiment twice. Accordingly the final sample consisted of 95 participants (80% female; 98% in educational training; 96% psychology students; mean age of *M* = 22.2 years [*SD* = 3.19]). Based on a one-tailed a priori power analysis for z-tests of two independent correlations with G*Power (alpha = .05, power = .8, rho1 = .2, rho2 = .0) we targeted 304 participants per condition, which however could not be achieved with the available resources. The participants were fairly evenly split between the two IATs, with 42 in the atypical IAT and 53 in the typical IAT.

Measures

IATs. The target categories of the typical as well as of the atypical IAT were again environmental protection/environmental degradation. The typical IAT consisted of five positive word exemplars for environmental protection (i.e., nature reserve, solar energy, fair trade, recycling, biodiversity) and five negative word exemplars for environmental degradation (i.e., deforestation, microplastics, waste pollution, overfishing, wasting resources). The atypical IAT consisted of five less positive word exemplars for environmental protection (i.e., radical environmental activists, airplane ban, consumption abstinence, climate protest, emissions tax) and five less negative word exemplars for environmental degradation (i.e., tropical timber, luxurious lifestyle, ski resorts, globalization, long-distance travel). We selected the target exemplars on the basis of the two pretests already described, using the same selection criteria as in Experiments 1 and 2 (see Supplement 3 for the specific descriptive statistics of the selected exemplars). For both IATs we used the same five positive exemplars for the attribute category good and the same five negative exemplars for the attribute category bad, which we selected from the attribute exemplars already used in Experiment 1. While we kept the modified IAT block structure from Experiment 2, the number of trials per block as compared to Experiment 2 changed due to the number of exemplars per category in Experiment 3 so that both IATs consisted of 13 blocks in which each exemplar stimulus of the given categories was presented once. Apart from these changes, the IAT procedure remained the same as in Experiments 1 and 2.

IAT effects were calculated and coded in the same way as in Experiments 1 and 2, so that positive D scores indicate a more positive evaluation of the target category environmental protection.

Outcome variables. We used the same outcome variables as in Experiments 1 and 2 (exemplar evaluation measure, target evaluation measure, and behavior measure with internal consistencies of $\omega_t = .83$, $\omega_t = .57$, and $\omega_t = .26$, respectively, in the typical IAT condition and of $\omega_t = .81$, $\omega_t = .74$, and $\omega_t = .55$, respectively, in the atypical IAT condition), the only difference being that the target exemplars of the exemplar evaluation measure corresponded to the selected target exemplars of the IAT in Experiment 3 (see the respective section of Experiment 1 for more details).

Data analysis

While the hypotheses were identical to those of Experiments 1 and 2, the change of the experimental design to a between design made it necessary to adjust our statistical analyses. We applied multigroup SEM (see Breitsohl, 2019; Ployhart & Oswald, 2004), with the two IAT conditions serving as experimental groups. This meant that this time, compared to Experiments 1 and 2, we were able to test H1-H3 in one model. The latent D score variable was measured via five indicators, i.e., five D-scores were calculated on the basis of the five test blocks of the IAT, and the outcome criteria were measured and modeled as in Experiments 1 and 2. A conceptual representation of the model fitted in both groups can be found in Figure 2 (with the only difference being that in Figure 2 the latent D score variable was measured via two instead of five indicators). We also controlled for potential block order effects, as in Experiments 1 and 2.

Results

Preliminary analyses

Ensuring data quality. We applied the same criteria for ensuring data quality as in Experiment 1 based on which 1.04% of the participants and 0.02% of the trials were excluded (refer to the corresponding section of Experiment 1 for a detailed description).

Descriptive statistics, multivariate normal distribution, handling missing values, and measurement invariance. Descriptive statistics for all manifest indicators can be found in Supplement 3 on our OSF project page. The manifest indicators for the latent multigroup model violated the assumptions of a multivariate normal distribution (Mardia's skewness = 409.94, $p < .001$; Mardia's kurtosis = 2.41, $p < .05$), which is why we used the MLMV estimator. There were no missing values for any of the manifest indicators. We first tested MI without including the outcome variables in the model and found that we could assume strong MI, since the chi square difference test between the weak and the strong MI model was non-significant, $\Delta\chi^2 = 3.12$, $p = 0.54$. Strong MI was needed to be able to interpret the latent mean and variance differences between the two IATs (Widaman & Reise, 1997). We then included the outcome variables and found that the strong MI model still had a very good model fit, S-B $\chi^2_{\text{strong MI}}$ (116) = 121.74, $p = .34$; RMSEA$_{\text{strong MI}}$ = 0.04; CFI$_{\text{strong MI}}$ = .95; SRMR$_{\text{strong MI}}$ = 0.10; AIC$_{\text{strong MI}}$ = 2241.8; BIC$_{\text{strong MI}}$ = 2395.1, and, accordingly, assumed strong MI for all of the following analyses.

Main analyses

H1: Test difficulty. As hypothesized, the latent mean of the atypical IAT, $\hat{\mu}_{\text{atypical}}$ = 0.77, was descriptively smaller than the latent mean of the typical IAT, $\hat{\mu}_{\text{typical}}$ = 1.02 (see Table 3 for the latent means and their standard errors). To test for significance, we added the group equality constraint that the latent means of the IATs were equal to the strong MI model. The resulting model (means model) showed a significantly worse fit compared to the strong MI model (see fit indices in Table 4), which indicates that the latent means of the IATs differed significantly. A Wald test further supported this result, $W(1) = 21.68$, $p < .001$. Furthermore, the IAT effect of the atypical IAT was still significantly different from zero, $z = 14.33$, $p < .001$.

Table 3

Latent Means, Latent True-Score Variances, Latent Correlations, and Reliabilities for the Typical and Atypical IAT in the Strong Invariance Multigroup Model (Experiment 3)

Condition	L mean (*SE*)	L variance (*SE*)	L cor target measure (CI)	L cor exemplar measure (CI)	L cor behavior measure (CI)	Reliability
Typical IAT	1.02 (.04)	0.03 (.01)	.36 (-.43, 1.08)	-.07 (-.83, .51)	.18 (-.27, .58)	.43
Atypical IAT	0.77 (.05)	0.03 (.02)	.04 (-.65, .60)	.08 (-.45, .73)	-.17 (-.57, .38)	.47

Note. L = latent; cor = correlation; CI = bootstrap-bias-corrected confidence intervals; IAT = implicit association test.

Table 4

Model Fit of the Different Models to Test the Overall Manipulation Hypotheses (Experiment 3)

Model	S-B χ^2 (df)	*p*	RMSEA	CFI	SRMR	AIC	BIC	$\Delta\chi^2$	*p*
Strong MI	121.74 (116)	.34	0.04	0.95	0.10	2241.8	2395.1		
Means	142.66 (120)	.08	0.08	0.79	0.14	2262.3	2405.3	28.95	<.001
Variances	124.67 (120)	.37	0.04	0.96	0.11	2236.3	2379.3	2.80	.59
Covariances	124.37 (119)	.35	0.04	0.95	0.11	2238.7	2384.2	2.56	.46

Note. S-B χ^2 = Satorra-Bentler scaled χ^2; RMSEA = robust root-mean-square error of approximation; CFI = robust comparative fit index; SRMR = robust standardized root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; MI = measurement invariance; Means = strong measurement invariance model plus equal group means; Variances = strong measurement invariance model plus equal group variances; Covariances = strong measurement invariance model plus equal group covariances.

H2: True-score variance. Contrary to our hypothesis, the true-score variance of the atypical IAT, $\widehat{\sigma}_{\text{atypical}}^2 = 0.03$, was identical to that of the typical IAT, $\widehat{\sigma}_{\text{typical}}^2 = 0.03$ (see Table 3 for the true-score variances and their standard errors). We nevertheless tested for significant differences for the sake of completeness. Accordingly, we added the group equality constraint that the true-score variances of the IATs were equal to the strong MI model. Not surprisingly, the resulting model (variances model) did not fit significantly worse than the strong MI model (see fit indices in Table 4), which indicates that the true-score variances of the IATs were not significantly different. This was further supported by a Wald test, $W(1) = 0.18$, $p = .67$.

H3: Predictive power. In contrast to our hypothesis, descriptively, the atypical IAT had smaller latent correlations with two of the three outcome variables than the typical IAT. The latent correlations of the atypical and the typical IAT with the target evaluation measure were $\hat{r}_{\text{atypical}} = .04$ and $\hat{r}_{\text{typical}} = .36$, respectively, with the exemplar evaluation measure they were $\hat{r}_{\text{atypical}} = .08$ and $\hat{r}_{\text{typical}} = -.07$, respectively, and with the behavior measure they were $\hat{r}_{\text{atypical}} = -.17$ and $\hat{r}_{\text{typical}} = .18$, respectively (see Table 3 for the latent correlations and CIs). Although the difference in correlations points in the wrong direction in two cases and is very small in the other case, we tested the differences for significance. Accordingly, we added the group equality constraint that the latent covariances of the IATs were equal to the strong MI model. The resulting model (covariances model) did not show a significantly worse fit than the strong MI model (see fit indices in Table 4), which indicates that the latent covariances between the IATs and the outcome variables were not significantly different for the two IATs. Since neither the latent variances nor the latent covariances of the IATs differed, it can be assumed that the latent correlations were equal as well. This was further supported by Wald tests: $W(1) = 0.66$, $p = .42$, for the target evaluation measure; $W(1) = 0.16$, $p = .69$, for the exemplar evaluation measure; $W(1) = 1.81$, $p = .18$ for the behavior measure.

Discussion

In Experiment 3, we manipulated the valence of the target exemplar stimuli in a between design and developed two Environmental protection/Environmental degradation attitude IATs, a typical IAT consisting of typically valenced target exemplars, and an atypical IAT consisting of atypically valenced target exemplars. As hypothesized the atypical compared to the typical IAT (a) had a less extreme IAT test difficulty, i.e., had a smaller average IAT effect, but contrary to our hypotheses, this more moderate test difficulty was not accompanied by (b) an increase in true-score variance or (c) an increase in predictive power as assessed by the I-E or I-C correlations. Note that the effect of the manipulation on IAT test difficulty was similarly strong as in Experiment 2 and that thus a between design does not seem to lead to a stronger effect of the typicality manipulation than a within design. It therefore cannot be ruled out that this effect was still not strong enough or that that the test difficulty was still too far away from moderate test difficulty (the IAT effect of the atypical IAT was not only very large, but in fact even larger than those of the atypical stimulus sets in Experiments 1 and 2) for positive downstream effects on true-score variance or predictive power to occur (see the General Discussion for a more detailed analysis on what the results imply for the test difficulty account).

General Discussion

In three experiments, we attempted to positively influence the test difficulty, true-score variance, and predictive power of attitude IATs with target categories that can be regarded a priori as generally positive or negative. For these IATs with clearly valenced targets, we aimed to reduce their extreme test difficulties (and thus shift their test difficulties more in the direction of moderate test difficulties), by manipulating the target exemplar valence from typically valenced (general valence of exemplars and target categories match) to atypically valenced (general valence of exemplars and target categories deviate such that the valence of the exemplars is less in the direction of the valence of the respective target categories) in order to increase their true-score variance and predictive power. In Experiment 1, we developed an IAT with the target categories environmental protection/environmental degradation, using typically valenced target exemplars (e.g., biodiversity/deforestation for environmental protection/environmental degradation) as well as atypically valenced target exemplars (e.g., climate tax/globalization for environmental protection/environmental degradation). In Experiment 2, we again developed an IAT with the target categories environmental protection/environmental degradation, but changed the IAT procedure slightly by increasing the number of test blocks and exchanging some of the typically as well as atypically valenced exemplars for both of the two categories. In Experiment 3, we used the same target categories once more and retained the modified number of test blocks from Experiment 2, but again exchanged some of the target exemplars and developed two IATs, one typical and one atypical IAT, consisting only of typically and atypically valenced target exemplars, respectively.

In all three experiments, consistent with our hypotheses, the use of atypically instead of typically valenced target exemplars led to significantly less extreme test difficulties, though the test difficulties were still far from being moderate, and accordingly influenced the average IAT-effect, which is in line with previous research (e.g., Bluemke & Friese, 2006; Gast & Rothermund, 2010; Govan & Williams, 2004). In contrast to our hypotheses, however, in all three experiments this significant difference in test difficulty did not result in a significant difference in either the true-score variance or the predictive power of the IATs.

In the following, we discuss how the results can be explained in the context of the test difficulty account and what significance they have for the account, describe implications of the results for the selection of target exemplars and relate them to previous recommendations, discuss implications of the results for the IAT as a diagnostic tool for measuring individual differences in attitudes, discuss limitations of our experiments, and end with a final conclusion.

Explanation and significance of the results with regard to the test difficulty account

Since the manipulation of the target exemplar valence from typically to atypically valenced exemplars led to less extreme test difficulties, but not to positive downstream effects on the true-score variance and the predictive power of the IATs, the question arises as to what the reasons for this could be. One possible explanation is that, on the one hand, the influence on the test difficulty by manipulating the valence of the target exemplars was not strong enough and, on the other hand, that this too small influence led to a test difficulty that was still too extreme and too far away from a moderate test difficulty in the case of the atypical stimulus exemplars. In Experiment 1, the IAT effect for the atypical stimulus set was 0.72 and for the typical stimulus set 0.81, in Experiment 2 the IAT effect was 0.66 and 0.88, respectively, and in Experiment 3 the IAT effect was 0.77 for the atypical IAT and 1.02 for the typical IAT. For comparison, in the experiment by Urban et al. (2024, study 3), in which effects of different reference categories were found not only for test difficulty but also for true-score variance and predictive power of the IATs, the average IAT effects of the IATs with a reference category that was opposite in valence to the category of interest were 1.0 and 0.81, whereas the average IAT effect of an IAT with a reference category that had a similar valence as the category of interest was -0.1. The effect of the manipulation was therefore much stronger and accordingly resulted in an IAT that clearly had a moderate test difficulty in the case of a reference category that matched the valence of the category of interest. Manipulating the valence of the target exemplars may be an inappropriate strategy to influence the true-score variance and predictive power of IATs if the effect of the manipulation on the test difficulty is only small. It should be noted, though, that we could not fully exploit the potential power of the manipulation of target exemplars in our study, with the target categories that were investigated (environmental protection/degradation), because no representative (as assessed via self-reports) atypical exemplars with opposite valence to their respective target categories could be found. This situation, however, may be characteristic for target categories that have a clear valence.

Another possible explanation for our failure to find evidence for the hypothesized effects is that atypical exemplars might have a detrimental influence on the attitude-related variance of IAT effects and their predictive power. Thus, although atypical target exemplars lead to a less extreme test difficulty, this effect might be counteracted by their triggering additional processes that undermine their validity. We can only speculate, what these additional processes are (e.g., associations to other categories and topics), but our results could indicate that atypical exemplars elicit such additional processes which might cancel out the positive effects of a less extreme test difficulty on true-score variance and predictive power.

For these reasons, we think that one should not necessarily jump to the conclusion that the results reported here invalidate the test difficulty account of the IAT. However, this possibility cannot be ruled out either. Urban et al. (2025) investigated the effects of using bivalent vs. univalent attributes in attitude IATs and found, similar to the present study, the hypothesized influence of the respective manipulation on IAT test difficulty, but no corresponding effects on true-score variance and predictive power. Thus, although there are plausible explanations for the lack of effects in both the present study and the study from Urban et al. (2025; e.g., less efficient manipulations of test difficulty), the results nonetheless call for further examination of the test difficulty account. Rather than being directly translated into true-score variance and predictive power, additional conditions could have to be met in order to translate test difficulty into true-score variance and relations to outcome variables (e.g., simple IAT task procedures, intuitive category labels, non-ambiguous exemplars).

Selecting target exemplars based on their valence: previous recommendations versus the effects of target exemplar valence on the psychometric properties of attitude IATs

There are two central recommendations for the selection of target exemplars (e.g., Greenwald et al., 2022; Teige-Mocigemba et al., 2010): 1) the exemplars of one target category should differ from the exemplars of the other target category only by one central feature, the nominal or semantic meaning that relates them to the respective category, and 2) the exemplars should be easy to categorize or representative of their respective target category. It follows from the first recommendation that, for example, the valence of the stimuli should not be positive for one target category and negative for the other, and that therefore, in the case of attitude IATs with clearly positive or negative target categories, not only typically valenced exemplars should be used. Instead, either positive and negative target exemplars should be balanced within each target category (e.g., Greenwald et al., 2022; Teige-Mocigemba et al., 2010) or neutral exemplars (e.g., Gast & Rothermund, 2010; Greenwald et al., 2022) or synonyms of the target categories should be used (Steffens et al., 2008). What the second recommendation means in practice is less clear. Greenwald et al. (2022) conclude that participants should be able to categorize the exemplars in the target discrimination practice block quickly (mean reaction time in the range of 600 to 800 ms) and with few errors (less than 10% error rate). A rather obvious problem of this conclusion is that the reaction time and error rate in the target discrimination practice block depend not only on the representativeness of the exemplars, but also on whether or not the valence of the exemplars matches the valence of the respective target categories, so that representativeness and valence are intertwined. Applying the second recommendation could therefore inadvertently lead to a violation of the first recommendation by selecting exemplars whose valence is confounded with the valence of the respective target categories. In fact, Greenwald et al. (2022) even seem to fall prey to this problem themselves, as they characterize atypically valenced stimuli as inherently non-representative. Given these potentially conflicting practical interpretations of the two recommendations, including the rather arbitrary criteria proclaimed by Greenwald et al. (2022) to assess the representativeness of the exemplars (no reasons are given for the chosen response times or error rates), it is not surprising that Hogenboom et al. (2024) conclude that, applying Greenwald et al.'s criteria, more than 94% of the 923 target and attribute exemplars they analyzed using large datasets from project implicit would have to be discarded.

Taken together, we therefore think that in the future it would be helpful to use psychometric properties describing the reliability and validity of the IAT as objective and valid criteria for exemplar stimulus selection (and that different features of the exemplars should be examined for their effect on these psychometric properties). Exemplars that improve the psychometric properties of the IAT compared to exemplars that worsen the psychometric properties can arguably be described as the more appropriate exemplars that should be selected. Ideally, these exemplars should then also fulfil the general recommendations described above, and if not, the recommendations should be revised.

In our case, considering the psychometric properties under investigation, that is, the test difficulty, true-score variance, and predictive power, results suggest that using atypically compared to typically valenced target exemplars influences the test difficulty, but none of the properties that are directly related to the reliability or validity of the IATs. This indicates that whether atypically or typically valenced exemplars are used might be less relevant than previously thought. However, these results must also be treated with caution as we will explain in more detail in the Limitations section.

Insights from the test difficulty account for the IAT as a measurement instrument for individual differences in attitudes

Our results are in line with a number of studies that question whether the IAT is a suitable method for measuring individual differences in attitudes (Payne et al., 2017; Schimmack, 2021a). In all three experiments, regardless of the exemplar stimuli considered, a) the true-score variance of the IATs was low ($0.03 \leq \widehat{\sigma}^2 \leq 0.06$), b) the reliability of the IATs was limited ($.39 \leq \omega_t \leq .48$), and c) the predictive power of the IATs was low no matter which outcome variable was examined ($-.33 \leq , \hat{r} \leq .36$; note that none of the correlations were significantly different from 0 in the assumed direction and that the confidence intervals were very large due to the low levels of true-score variance).

These results are not only sobering with regard to the general quality of the IAT to measure individual differences in attitudes, here specifically in environmental attitudes, they also point to another problem, namely that the psychometric properties of IATs cannot be easily influenced by changing the design of IATs. In our case manipulating the valence of the exemplar stimuli had no effect on the true-score variance and the predictive power of the IATs. Other recent results from the test difficulty account go in a similar direction. Urban et al. (2025) found that the use of bivalent vs. univalent attribute categories had no effect on the true-score variance and, in most cases, no effect on the predictive power of the IATs (only one of 15 correlation comparisons was significant). This is not to say, of course, that it is impossible to influence the psychometric properties of IATs. Within the test difficulty account Urban et al. (2024, study 3) provided first evidence that manipulating the valence of the reference category while keeping the relevant target category constant can influence both the true-score variance and the predictive power of IATs. Nevertheless, our results demonstrate how difficult it is to influence the psychometric properties of IATs, which is also supported by studies outside the test difficulty account; for example, studies in which attribute exemplar characteristics were manipulated and no effects on psychometric properties were found (Axt et al., 2021; Stieger et al., 2010). If the usual recommendations for developing an IAT are followed (cf. Greenwald et al., 2022) and the psychometric properties are still unsatisfactory, the adaptations that IAT researchers can make to the IAT are limited, given the interest in specific attitude objects.

In comparison, questionnaires measuring individual differences in attitudes appear to be much easier to influence in terms of their psychometric properties: The number of scale points and the labelling of the scale points (e.g., Alwin & Krosnick, 1991; Krosnick &

Fabrigar, 1997), the valence of the item formulation (Chang, 1995a, 1995b), and, above all, the content of the item, all these and many other characteristics influence the psychometric properties of the respective measurement instrument. To summarize, our results are consistent with Corneille and Gawronski's (2024) recent and compelling demonstration that IATs are inferior to questionnaires as measurement instruments for individual differences. Note, however, that our critique relates only to the measurement of individual differences in attitudes and does not necessarily apply to the measurement of individual differences in other constructs, such as wanting or stereotypes, or to other purposes for which the IAT is used, such as testing personality theories (Schimmack, 2021a) or measuring situations (Payne et al., 2017).

Limitations

A potential limitation of our experiments is that, as already described, we only focused on one particular attitude domain, namely environmental protection/environmental degradation. While our results confirm that these target categories generally elicit positive or negative evaluations – a necessary condition for applying our manipulation – we could not find any representative (as assessed by self-reports) atypical exemplars with opposite valence to their respective target categories, nor could we find enough representative atypical exemplars with neutral valence such that the atypical stimulus sets or the atypical IAT could consist exclusively of neutral exemplars.

With regard to the test difficulty account, this means that we may be underestimating the effect of our manipulation. To circumvent this problem, it is possible on the one hand to relax the criterion of representativeness, since, as already discussed, it is unclear how the representativeness of the exemplars can be empirically determined, and on the other hand it is possible to think of other attitude domains with clearly positive or negative categories for which it might be easier to find representative exemplars with opposite valence to their respective target categories or a sufficient number of neutral exemplars (e.g., old vs. young).

Under these circumstances, the approach of manipulating the valence of the target examples could lead to a stronger effect than was the case in our experiments, and thus also lead to the hypothesized effects. Choosing exemplars with opposite valence to their target categories, however, will result in sets of exemplars that are non-balanced in terms of their valence (e.g., positive exemplars for the category old vs. negative exemplars for the category young), which would also violate recommendations for using balanced sets of exemplars. Then again, the objective of our study was to shift IATs with extreme test difficulty to a less extreme test difficulty, and the more extreme the test difficulty, the more likely it is that such exemplars will be difficult to find even under a relaxed criterion of representativeness, so that we may underestimate the effect of the manipulation, but this would not change the fact that the manipulation would still not be the best approach to influence IAT test difficulty, true-score variance, and predictive power due to the difficulty of practical implementation. With regard to the selection of target exemplars we cannot exclude the possibility that the exclusive use of exemplars with neutral valence or the counterbalancing of atypical and typical exemplars may have an influence on the psychometric properties of the IAT and that target exemplar valence may therefore be more important for the development of IATs than our results suggest.

Another problem that could go hand in hand with focusing on the particular attitude domain environmental protection/environmental degradation is that the true-score variance and the I-E or I-C correlations within this domain might be inherently too low for the effect of the manipulation to emerge. We may have inadvertently chosen an inappropriate attitude domain to test our hypotheses, and another attitude domain which in principle does allow for a greater variety in true-score variance and correlations between IATs and outcome variables may have produced the hypothesized results. However, this alternative explanation seems unlikely, the main reason being that it has already been shown that IATs can be developed in this attitude domain that correlate significantly with direct attitude measures similar to the target evaluation measure used in our experiments (Urban et al., 2024, study 3).

Conclusion

We attempted to address the long-standing problem of low predictive power of the IAT and modified the design of attitude IATs with target categories expected to generally evoke clearly positive or negative evaluations by changing the target exemplar stimuli from typically to atypically valenced ones. Based on the test difficulty account (Urban et al., 2024), we hypothesized that this manipulation should lead to a less extreme test difficulty and, consequently, to a higher true-score variance as well as to a higher predictive power of the resulting IAT scores. However, results from three experiments indicate that while atypically compared to typically valenced exemplars result in IATs with less extreme test difficulty, they do not result in increased true-score variance or predictive power. Accordingly, manipulating the valence of target exemplars does not seem to be a suitable strategy for positively influencing the psychometric criteria investigated. Conversely, however, our results seem to suggest that the valence of the target exemplars appears to be less relevant than perhaps previously assumed (cf. Bluemke & Friese, 2006; Greenwald et al., 2022; Teige-Mocigemba et al., 2010), at least with regard to the psychometric criteria we investigated.

References

Alwin, D. F., & Krosnick, J. A. (1991). The Reliability of Survey Attitude Measurement. *Sociological Methods & Research*, *20*(1), 139–181. https://doi.org/10.1177/0049124191020001005

Axt, J. R., Feng, T. Y., & Bar-Anan, Y. (2021). The good and the bad: Are some attribute words better than others in the Implicit Association Test? *Behavior Research Methods*, *53*(6), 2512–2527. https://doi.org/10.3758/s13428-021-01592-8

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT. *The Journal of Applied Psychology*, *94*(3), 567–582. https://doi.org/10.1037/a0014665

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, *42*(2), 163–176. https://doi.org/10.1016/j.jesp.2005.03.004

Breitsohl, H. (2019). Beyond ANOVA: An Introduction to Structural Equation Models for Experimental Designs. *Organizational Research Methods*, *22*(3), 649–677. https://doi.org/10.1177/1094428118754988

Chang, L. (1995a). Connotatively Consistent and Reversed Connotatively Inconsistent Items are Not Fully Equivalent: Generalizability Study. *Educational and Psychological Measurement*, *55*(6), 991–997. https://doi.org/10.1177/0013164495055006007

Chang, L. (1995b). Connotatively Inconsistent Test Items. *Applied Measurement in Education*, *8*(3), 199–209. https://doi.org/10.1207/s15324818ame0803_1

Corneille, O., & Gawronski, B (2024). Self-reports are better measurement instruments than implicit measures. *Nature Reviews Psychology*, *3*(12), 835–846. https://doi.org/10.1038/s44159-024-00376-z

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*(6), 443–451. https://doi.org/10.1006/jesp.2000.1464

Gast, A., & Rothermund, K. (2010). When old and frail is not the same: Dissociating category and stimulus effects in four implicit attitude measurement methods. *The Quarterly Journal of Experimental Psychology*, *63*(3), 479–498. https://doi.org/10.1080/17470210903049963

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. *Journal of Experimental Social Psychology*, *40*(3), 357–365. https://doi.org/10.1016/j.jesp.2003.07.002

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., . . . Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, *54*(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 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*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Hogenboom, S. A. M., Schulz, K., & van Maanen, L. (2024). Implicit association tests: Stimuli validation from participant responses. *British Journal of Social Psychology*, *63*(2), 975–1002. https://doi.org/10.1111/bjso.12688

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In L. E. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). Wiley. https://doi.org/10.1002/9781118490013.ch6

Kurdi, B., & Banaji, M. R. (2017). Reports of the Death of the Individual Difference Approach to Implicit Social Cognition May Be Greatly Exaggerated: A Commentary on Payne, Vuletich, and Lundberg. *Psychological Inquiry*, *28*(4), 281–287. https://doi.org/10.1080/1047840X.2017.1373555

Machery, E. (2022). Anomalies in implicit attitudes research. *WIREs Cognitive Science*, *13*(1), Article e1569. https://doi.org/10.1002/wcs.1569

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting Behavior With Implicit Measures: Disillusioning Findings, Reasonable Explanations, and Sophisticated Solutions. *Frontiers in Psychology*, *10*, 2483. https://doi.org/10.3389/fpsyg.2019.02483

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*, *104*(1), 45–69. https://doi.org/10.1037/a0030734

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

Payne, B. K, Vuletich, H. A., & Lundberg, K. B. (2017). The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. *Psychological Inquiry*, *28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Ployhart, R. E., & Oswald, F. L. (2004). Applications of Mean and Covariance Structure Analysis: Integrating Correlational and Experimental Approaches. *Organizational Research Methods*, *7*(1), 27–65. https://doi.org/10.1177/1094428103259554

R Core Team (2021). R: A language and environment for statistical computing. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schimmack, U. (2021a). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Schimmack, U. (2021b). Invalid Claims About the Validity of Implicit Association Tests by Prisoners of the Implicit Social-Cognition Paradigm. *Perspectives on Psychological Science*, *16*(2), 435–442. https://doi.org/10.1177/1745691621991860

Steffens, M. C., Kirschbaum, M., & Glados, P. (2008). Avoiding stimulus confounds in Implicit Association Tests by using the concepts as stimuli. *The British Journal of Social Psychology*, *47*(2), 217–243. https://doi.org/10.1348/014466607X226998

Stieger, S., Göritz, A. S., & Burger, C. (2010). Personalizing the IAT and the SC-IAT: Impact of idiographic stimulus selection in the measurement of implicit anxiety. *Personality and Individual Differences*, *48*(8), 940–944. https://doi.org/10.1016/j.paid.2010.02.027

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A Practical Guide to Implicit Association Tests and Related Tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (1st ed., pp. 117–139). The Guilford Press. https://escholarship.org/content/qt63t6n75d/qt63t6n75d.pdf

Urban, M., Koch, T., & Rothermund, K. (2024). The Implicit Association Test and its difficulty(ies): Introducing the test difficulty concept to increase the true-score variance and, consequently, the predictive power of implicit association tests. *Journal of Personality and Social Psychology*, *127*(1), 31–57. https://doi.org/10.1037/pspa0000391

Urban, M., Koch, T., & Rothermund, K. (2025). Effects of Bivalent Versus Univalent Attribute Categories on Test Difficulty, True-Score Variance, and Predictive Power of Attitude Implicit Association Tests. *Collabra: Psychology*.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant (Ed.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

CRediT author statement

xxx: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing- Original draft. **xxx:** Formal analysis, Methodology, Resources, Validation, Writing—review and editing. **xxx:** Conceptualization, Funding acquisition, Methodology, Resources**,** Supervision, Writing—review and editing.

Ethics statement

The research received ethical approval from the ethics committee of the University of xxx (FSV 24/021) and was in accordance with the Declaration of Helsinki as amended in 2013.

Competing interests

We have no competing interests to disclose.

Data accessibility statement

All study material, data, codebooks, and code, as well as the preregistrations of our hypotheses and analyses plans, are publicly available at our project's Open Science Framework page (Link: https://osf.io/3meza/?view_only=dcdfd55cc56a45d6929c3d391dd0ce2c).

List of Figures