Effects of Bivalent Versus Univalent Attribute Categories on Test Difficulty, True-Score Variance, and Predictive Power of Attitude Implicit Association Tests

Merlin Urban, Tobias Koch, & Klaus Rothermund

FSU Jena, Germany

Collabra: Psychology, in press

Abstract

Based on the test difficulty account, we manipulated the attribute categories of Implicit Association Tests (IATs), using either Bivalent (e.g., good/bad) or Univalent (e.g., good/very good) evaluative adjective pairs. To increase the true-score variance and the predictive power of IATs, we sought to shift their test difficulty from extreme test difficulty (in case of Bivalent IATs) to moderate test difficulty (in case of Univalent IATs). In Experiment 1 (n = 193) we developed a Bivalent and a positive Univalent single-target IAT. In Experiment 2 (n = 180) we developed a Bivalent, a positive Univalent and a negative Univalent standard IAT. In both experiments, the Univalent IATs were significantly closer to moderate test difficulty than the Bivalent IATs, but did not show significantly more true-score variance or more predictive power. Based on our results, we would advise IAT researchers against using Univalent evaluative adjective pairs as attribute categories in the future.

Keywords: Implicit Association Test (IAT), test difficulty, predictive power, attribute categories, Bivalent/Univalent evaluative adjectives

Recently the concept of test difficulty has been introduced to research on the IAT (Urban et al., 2024) to tackle the long-standing issue of low predictive power of the IAT, that is, the insufficient ability of the IAT to predict relevant outcome variables (Blanton et al., 2009; Meissner et al., 2019; Meissner & Rothermund, 2025; Oswald et al., 2013; see also Corneille & Gawronski, 2024, and Gawronski & Corneille, 2025, for a more general criticism of the limited incremental predictive power of implicit measures). Based on the test difficulty account it was shown that, in accordance 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 difficulties.

Urban et al. (2024) defined IAT test difficulty in a technical sense, drawing on the test difficulty concept of CTT: Accordingly, IAT test difficulty indicates to what extent people answer in the keyed direction of the theoretical construct. For example, in the case of an attitude IAT, IAT test difficulty indicates to what extent people answer in favor of the attitude construct. The mean test score, that is, the average IAT effect, indicates the IAT test difficulty.[1] Due to the composition of the IAT (its relativity and block structure), two conditions must be met to interpret the average IAT effect in terms of IAT test difficulty: 1) It must be defined which target category serves as attitude object in which keyed direction it is answered or not (called *relevant target category*) and 2) It must be defined in which block this target category shares a response key with the attribute category expressing the keyed direction (e.g., the positive attribute in a typical attitude IAT). Suppose an environmental protection/environmental degradation attitude IAT. Say we define environmental protection to be the relevant target category and the block in which the relevant target category and positive share a response key as the subtrahend in the

[1] Importantly, IAT test difficulty does not refer to how difficult it is to perform the task ("task difficulty"), that is, conceptually it is unrelated to the average response time or error frequency of an IAT.

calculation of the IAT effect. Then a large positive average IAT effect would indicate that the IAT is an easy test because participants would have answered strongly in the keyed direction of the theoretical construct and evaluated environmental protection more positively then environmental degradation (that is, reacted faster on average when the relevant target category "environmental protection" and the attribute "positive" were assigned to the same response key). A large negative average IAT effect would indicate that the IAT is a difficult test because participants would have answered strongly in opposition to the keyed direction of the theoretical construct and evaluated environmental protection less positively then environmental degradation (that is, reacted slower on average when the relevant target category "environmental protection" and the attribute "positive" were assigned to the same response key). Finally, an average IAT effect of (or close to) zero would indicate that the IAT has a moderate difficulty because participants would have neither answered more strongly in the keyed direction nor in the opposite direction (that is, neither reacted faster nor slower on average when the relevant target category "environmental protection" and the attribute "positive" were assigned to the same response key).[2]

Based on the test difficulty account, Urban et al. (2024) derived three approaches to modify the IAT design in order to establish IATs of moderate test difficulty: Manipulating the valence of the target reference category, manipulating the valence of the attribute categories, and manipulating the valence of the exemplars of the target categories. They provided evidence that the test difficulty of an IAT can be influenced towards moderate test difficulty by choosing a reference category (in their case leisure time) that has a similar valence to the relevant target category (in their case environmental protection), and could show that this has positive

[2] Empirically speaking, in the given example of an environmental protection/environmental degradation attitude IAT, in which we define "environmental protection" to be the relevant target category, the IAT would be a very easy test.

downstream effects on true-score variance and predictive power of the resulting IAT (Urban et al., 2024, study 3).

In this study, we will explore the second of these approaches in more detail. Specifically, we manipulated the valence of the attribute categories of attitude IATs, by comparing IATs that employ Bivalent evaluative adjective pairs (i.e., adjectives with opposite valences [bad vs. good]) with IATs that employ Univalent evaluative adjective pairs (i.e., adjectives denoting different grades of valence extremity within a single valence category [positive vs. extremely positive, or negative vs. extremely negative]).

The idea behind manipulating the attribute categories of attitude IATs to influence IAT test difficulty, true-score variance and predictive power

In the context of attitude questionnaires, Bivalent evaluative adjective pairs are commonly used as scale anchors when assessing attitude objects that are assumed to elicit a wide range of evaluative responses, ranging from negative to positive (e.g., abortion, Obamacare). This ensures that the full range and variability of possible evaluations is optimally captured. However, when assessing attitude objects that are assumed to be uniformly evaluated as either positive (e.g., peace, environmental protection) or negative (e.g., war, pollution), it may be more promising to use Univalent evaluative adjective pairs to optimally capture the variability within each valence category.

To illustrate, consider an attitude object that is expected to elicit only positive evaluations, such as environmental protection, and that is measured on a 7-point Likert scale with the Bivalent adjective pair negative vs. positive as scale anchors. In this case, it can be assumed that most responses would be above the midpoint of the scale, leaving the lower half unused. This ceiling effect would be accompanied by the fact that the item is easy, and its ability to discriminate between individuals (true-score variance) as well as to predict outcome variables (predictive

power) would be compromised. If the same attitude object is measured on a 7-point Likert scale with the Univalent adjective pair positive vs. extremely positive, the entire scale might be used, thereby reducing the ceiling effect. In other words, the Univalent scale should differentiate better between individuals who differ with regard to the degree of endorsing or advocating environmental protection. This should be accompanied by a shift towards more moderate item difficulty, more true-score variance, and a higher predictive power.

Note that, to our knowledge, there are no empirical studies that have investigated the effects of Bivalent vs. Univalent evaluative adjective pairs on the measurement quality of items for attitude assessment as a function of attitude object valence. Although some studies have looked at the effect of using different types of adjective pairs as anchors in rating scales (e.g., disagreement – agreement vs. no agreement – agreement; Höhne et al., 2021; Höhne et al., 2022; Menold, 2021; Menold & Raykov, 2016), none of the studies examined effects of evaluative adjective pairs (e.g., negative – positive vs. positive – extremely positive).[3]

In the context of attitude IATs, the evaluative adjective pairs, that is, the attribute categories, are typically Bivalent and are not individually matched to the attitude objects, that is, the target categories (cf. Greenwald et al., 2022). However, such a matching approach may also be promising in the context of IAT construction. The IAT is a binary classification task and as such, response tendencies for generally positive or generally negative attitude objects may be very homogeneous for Bivalent IATs, even across different degrees of attitude endorsement. Thus, attitude IATs with Bivalent attribute categories should result in IATs with extreme test difficulty, low true-score variance, and low predictive power; in contrast, changing the Bivalent

[3] Note that although the construction of questionnaires is not the main focus of this work, we also report initial exploratory results on this aspect in the General Discussion.

attribute categories to Univalent attribute categories should result in more moderate difficulties, more true-score variance, and more predictive power.

To better understand the rationale behind the idea in the IAT framework, let us consider a somewhat simplified situation, that is, a single-target (ST) IAT (Bluemke & Friese, 2008; Wigboldus et al., 2004) or a single-category (SC) IAT (Karpinski & Steinman, 2006). The advantage of considering an ST-IAT or an SC-IAT is that we eliminate the reference category (the non-relevant target category) to simplify matters, and to better understand the individual effect of the attribute categories, since the reference category might also affect IAT test difficulty, true-score variance, and predictive power (cf. Urban et al., 2024, study 3). Suppose we want to assess attitudes towards a relevant target category that is a priori considered to be generally positive (like environmental protection). If one now uses Bivalent attribute categories such as negative vs. positive, then this should result in a large positive average IAT effect, that is, a very easy IAT. Almost all people would respond faster in the block in which environmental protection and positive share the same response key than in the block in which environmental protection and negative share the same response key. However, if one changes the Bivalent attribute categories negative vs. positive to the Univalent attribute categories positive vs. extremely positive, then this should lead to a markedly weaker positive average IAT effect, that is, a less easy IAT. Some people will associate environmental protection with ‘positive’ (i.e., with the less positive attribute category of the pair) and respond faster when these two categories are paired on the same response key while others will associate environmental protection with ‘extremely positive’ (i.e., with the more positive attribute category of the pair) and respond faster when these two categories are paired on the same response key. If the IAT is calibrated so that the average valence of the target category lies between the two Univalent attribute categories, the IAT should

be closer to moderate difficulty and discriminate better between individuals who all endorse the target category, but to different degrees.

The same rationale should also apply to a standard IAT with two target categories of opposite valence despite the additional influence of the reference category. For example, if the relevant target category for an IAT is positive (like environmental protection), positive Univalent attribute categories should be used. The respective reference category would be negative (in this case environmental degradation) and should not influence responding differently in the two blocks of the IAT because its valence would not match any of the attribute categories. We thus would expect that the evaluations towards a relevant target category with an unambiguous valence are better captured when Univalent attribute categories are used than when Bivalent attribute categories are used, following the same reasoning described for the ST-IAT or the SC-IAT. The use of Univalent attribute categories for genuinely positive (and negative) target categories in the context of the standard IAT is all the more plausible if one considers that IAT researchers are often particularly interested in only one of the two target categories, such as the target category Black people in the case of a Black people vs. White people IAT. Accordingly, they are primarily interested in measuring differences in the evaluation of this one relevant target category with high accuracy.

Previous research on the valence/polarity of attribute categories in the context of IAT research

As far as we know the described ideas have not been systematically tested and empirically investigated yet. There are studies that refer to and examine the concept of polarity in the context of IAT research, but in none of these studies the focus is on Bivalent vs. Univalent attribute categories. Proctor and Cho (2006), proposed polarity correspondence as a general underlying principle in binary classification tasks, including the IAT. In their terminology they focused on

(the correspondence between) structural asymmetries between attributes, target categories, and responses, which are assumed to always share features of asymmetric polarities, which then produces S-R compatibility effects (Kornblum et al., 1990; Rothermund & Wentura, 2004). Following Pratkanis (1989), among others, Nosek (2005) argued that evaluations can have a Bipolar structure (i.e., attitude objects can be located on a Bipolar continuum, where liking one implies disliking the other, e.g., gun control vs. gun rights) or a Unipolar structure (i.e., attitude objects can be located on a Unipolar continuum, where liking one does not imply disliking the other, e.g., women vs. men). He found that the relationship between IATs and direct measures was stronger for evaluations with a Bipolar than a Unipolar structure. Kurdi et al. (2019) found that attribute categories with higher polarity (e.g., fat vs. thin) in comparison to attribute categories with lower polarity (e.g., sad vs. angry) had more predictive power. All of these investigations, however, are unrelated to the distinction between Bivalent and Univalent evaluative adjective pairs as attribute categories. The lack of studies relating to the ideas outlined, despite their theoretical soundness, indicates that a systematic empirical investigation of the effects of Univalent vs. Bivalent attributes on IAT effects is novel and promising.

Hypotheses and overview of the experiments

Our previous considerations can be summarized in the following hypotheses: Changing the attribute categories of an IAT with target categories that are a priori thought to be unambiguously positive or negative from Bivalent to Univalent attributes (a) shifts the test difficulty from a more extreme to a more moderate difficulty, that is, shifts the IAT effect from being strongly different from zero to being closer to zero, (b) increases the true-score variance, and (c) increases the predictive power of the resulting IAT.

We tested our hypotheses with both ST-IATs and standard IATs. In Experiment 1, we first used ST-IATs. As already illustrated, ST-IATs allow to test the hypothesized influence of

the valence of the attribute categories without the risk of unwanted influences of a reference category. However, there are other problems associated with the use of ST-IATs. For instance, it has been pointed out in the literature that its internal validity might be compromised due to undesirable focusing and recoding strategies resulting from the use of three instead of four categories (cf. Nosek et al., 2007; Rothermund & Wentura, 2010), and it has been found that its internal consistencies generally appear to be lower than those of the standard IAT (e.g., Axt et al., 2024; Bluemke & Friese, 2008; Karpinski & Steinman, 2006; but see also Greenwald & Lai, 2020, who found similar internal consistencies for the ST-IAT and the standard IAT). It is probably due to these reasons that the ST-IAT has received less attention as well as application (e.g., Kurdi et al., 2019), although it should be noted that there is recent literature indicating a better performance of the ST-IAT compared to the standard IAT on other psychometric criteria (e.g., Axt et al., 2024). Nevertheless, using only ST-IATs to test our hypotheses might not only be accompanied by the described psychometrical issues, but would also reduce the applicability of our results. Thus, in Experiment 2, we additionally tested our hypotheses using a standard version of the IAT with two target categories. The problems of the ST-IAT are thereby eliminated, although the aforementioned possibility of the reference category producing unintended influences remains. Since both IAT variants are accompanied by different advantages and risks, it seems to be the best and most comprehensive approach to test our hypotheses in the context of both IAT variants.

Furthermore, we explicitly took care to adhere to Greenwald et al.'s (2022) criteria for the construction of IATs in both our experiments. In order to also fulfil the criterion that the stimulus exemplars should be easy to sort, which might seem questionable in the case of Univalent attribute categories, we took additional precautions: Most importantly, in both experiments, we only used exemplars that a) were synonyms for the respective attribute categories and that b)

were very similar in valence to the attribute categories (note that the valence of the attribute categories clearly differed). A more detailed description of the selection of the exemplar stimuli including other smaller adjustments to ensure that the exemplars are easy to sort can be found in the Measures sections of Experiments 1 and 2.

To evaluate the predictive power of the IATs, we report the results of direct attitude measures in the form of gut reactions and actual feelings towards the target categories as outcome variables in both experiments (that is, we compared implicit-explicit [I-E] correlations), ensuring comparability with previous results of the test difficulty account (cf. Urban et al., 2024). However, to investigate the generalizability of our results, we also collected additional outcome variables: a) a behavioral measure in Experiment 1, so that we could also compare implicit-criterion (I-C) correlations, and b) two additional direct attitude measures in Experiment 2.

Experiment 1

In Experiment 1, we developed two ST-IATs with the same target category that was expected to elicit generally positive evaluations. As such target category we chose "environmental protection". The two ST-IATs were a Bivalent ST-IAT with the attribute categories bad vs. good (from now on called Bivalent ST-IAT) and a Univalent ST-IAT with the positive attribute categories good vs. very good (from now on called UnivalentPos ST-IAT).

Methods

Design and Procedure

We used a mixed design: The within factor *attribute category* had two levels (Bivalent attribute categories bad/good vs. Univalent positive attribute categories good/very good), the between factor *IAT order* had two levels (Bivalent ST-IAT first vs. UnivalentPos ST-IAT first), and the between factor *block order* had two levels (the compatible vs. incompatible block of the IAT was presented first). Participants were randomly assigned to the different conditions. They

were then asked to provide their demographic data and to complete the first ST-IAT. After that a questionnaire followed. In a final step they were asked to complete the second ST-IAT. Placing the questionnaire between the two ST-IATs allowed us to analyze our data also in terms of a between design by removing the second ST-IAT from all analyses (in case of significant interactions with IAT order). We preregistered this possibility and also implemented it in the reported analyses. We opted for the between design mainly because of two reasons: first, we found a near-significant interaction between the factors attribute category and IAT order (see Supplement 1 on our OSF project page for results), and second, it provides better comparability with the results of Experiment 2, which was developed as a between design from the start.

Sample

A total of $n = 196$ participants took part in Experiment 1. Participants were recruited via mailing lists of the University of xxx and via social networks. A prerequisite for participation in the study was that the participants' native language was German. Students from the University of xxx received course credits. Three participants had to be discarded for data quality reasons (see the Results section for a more detailed overview of the exclusion criteria) resulting in a final sample of 193 participants (75% female; 59% in educational training; mean age of $M = 31.36$ years [$SD = 15.41$]). Participants were distributed fairly evenly across the two ST-IATs (Bivalent ST-IAT: $n = 107$; UnivalentPos ST-IAT: $n = 86$). The final sample slightly exceeded the targeted 185 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 = .1, rho2 = .3, rho3 = .4). Note that this a priori power analysis was based on the initial within design and not

the between design that was ultimately utilized and that the expected effect sizes were deduced from Kurdi and Banaji (2019).[4]

Measures

ST-IATs. The two ST-IATs only differed with regard to the attribute categories and the corresponding exemplar stimuli. The exemplars for the attribute categories bad vs. good of the Bivalent ST-IAT were five negative (i.e., lousy, nasty, flawed, negative, awful) and five positive (i.e., pleasant, fine, neat, positive, beautiful) adjectives; the exemplars for the attribute categories good vs. very good of the UnivalentPos ST-IAT were the same five positive adjectives as for the attribute category good in the Bivalent ST-IAT and five very positive (i.e., ingenious, brilliant, phenomenal, grand, outstanding) adjectives. In addition to the typical recommendations for the selection of exemplars (e.g., Greenwald et al., 2022) we employed the following criteria to make sure that the adjectives could be categorized to the respective attribute categories unambiguously: we chose them a) to be similar in meaning as the respective attribute categories based on synonym dictionaries and b) to be similar in valence as the respective attribute category based on a pretest in which a total of 30 participants rated 64 evaluative adjectives on a 11-point Bivalent scale with endpoints ranging from 0 (*very negative*) to 10 (*very positive*).[5]

The remaining description of the ST-IATs applies to both ST-IATs. The exemplars for the target category environmental protection were five words similar in meaning to the target category (i.e., ecological, climate protection, sustainable, environmentally concerned, nature conservation). The exemplars were again selected based on the typical recommendations (e.g., Greenwald et al., 2022) as well as on a) synonym dictionaries and b) the previously described

[4] A one-tailed sensitivity power analysis for z-tests of two independent correlations with G*Power (alpha = .05, power = .8, n_1 = 107, n_2 = 86) showed that we could detect an effect size of q = .37. Note that the sensitivity analysis is only a conservative estimate of the true sensitivity of our multigroup structural equation models.

[5] The adjective "positive" was not included in the pretest, but appeared to be more suitable for selection than the other adjectives that were part of the pretest. Note also that we included the adjective "positive" in the pretest for the second experiment and, unsurprisingly, found a good match with the attribute category "good".

pretest, whereby this time we asked participants first to name synonyms for the target category in an open-ended question and second to indicate on a 7-point scale with endpoints ranging from 1 (*totally disagree*) to 7 (*totally agree*) how much they agreed that the words presented were synonyms for the target category.[6] The ST-IATs consisted of five blocks: the attribute discrimination practice block (20 trials), a short initial combined test block (20 trials), a long initial combined test block (40 trials), a short reversed combined test block (20 trials), and a long reversed combined test block (40 trials). Consequently, we have slightly modified the block structure compared to Bluemke and Friese (2008) as was already done elsewhere (e.g., Raccuia, 2016) with the goal to counteract the low reliability of ST-IATs reported in the literature. Since in the combined blocks of the ST-IAT two categories are always paired and assigned to a shared response key, whereas one attribute category is not paired and is the only category that is assigned to the other key, the exemplars of the unpaired attribute category were presented twice as often as those of the other two categories, to make sure that the left and right response keys were used equally often. Consequently, the exemplars of the target category, the paired attribute category, and the unpaired attribute category occurred in the short combined test blocks in a ratio of 5:5:10 trials and in the long combined test blocks in a ratio of 10:10:20 trials. The presentation of the attribute and target exemplars was randomized within each block. They were displayed against a white background and participants were instructed to respond to the exemplars as quickly and accurately as possible by pressing the left (*E*) or right (*I*) response key. If they did not respond within three seconds after the stimulus presentation or made an incorrect categorization, a corresponding feedback message was displayed in red font at the bottom of the page. In this case, participants were instructed to proceed with the correct response key.

[6] The word “nature conservation” was not included in the pretest. Instead the word “nature protection” was included and also often named by the participants in the open question which we replaced with “nature conservation” to avoid having two exemplar stimuli ending with “protection”.

IAT effects were calculated based on the D score algorithm (Greenwald et al., 2003).[7] D scores were calculated in such a way that positive D scores indicate a more positive evaluation of the target category environmental protection.

Outcome variables. We measured gut reactions and actual feelings towards the target category using questionnaire items. The items were: (a) "Rate your gut reactions towards environmental protection", and (b) "Rate your actual feelings towards environmental protection". Both items were to be rated on a 10-point Univalent scale with endpoints ranging from 1 (*slightly positive*) to 10 (*very positive*). On top of this central outcome variable, we collected two additional outcome variables. Firstly, Participants also had to rate both of the described items on a 10-point Bivalent scale with endpoints ranging from 1 (*very negative*) to 10 (*very positive*). Secondly, we collected a measure of environmentally friendly behavior using the General Ecological Behavior Scale (GEB; Kaiser & Wilson, 2004).

Data analysis

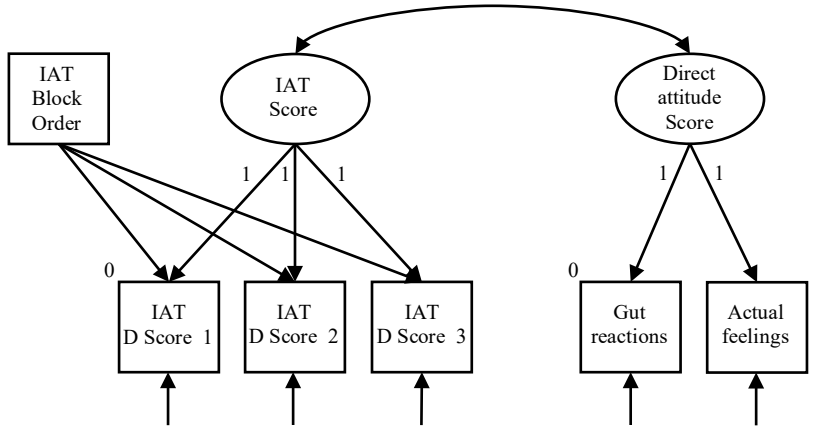
To test our hypotheses, we applied multigroup structural equation modeling (SEM; see Breitsohl, 2019; Ployhart & Oswald, 2004), with the two ST-IATs as experimental groups. This approach allowed us to test all our hypotheses in a single, unified statistical framework. Specifically, we were able to examine whether the two ST-IATs differed in terms of their a) latent means to establish whether manipulating the attribute categories resulted in the expected IAT test difficulties, b) latent variances to establish whether possible differences in IAT test difficulties would be associated with the expected IAT true-score variances, and c) latent correlations to establish whether possible differences in IAT test difficulty and IAT true-score variance yielded the expected differences in predictive power. The model fitted in all groups is

[7] We accounted for the different number of trials per category in the combined blocks by first calculating the respective estimates for the individual categories before averaging across the categories.

shown in Figure 1. The model consisted of two correlated latent variables. The latent D score variable was measured via three indicators, i.e., three D scores that were created via parcels at the trial level. The latent direct attitude variable was measured via two indicators, i.e., the Univalent gut reactions and the Univalent actual feelings.[8] Additionally, we controlled for potential block order effects by regressing the manifest IAT variables on the manifest covariate IAT block order. The manifest covariate was centered at the grand mean before the analysis.

Figure 1

Basic correlated two factor model that was fitted in both groups



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

Transparency and Openness

The research study received ethical approval from the ethics committee of the University of xxx and was in accordance with the Declaration of Helsinki as amended in 2013. Participants gave written informed consent to participate in this study. All study-related materials, including preregistration, stimulus material, raw data, curated data, codebooks, data cleaning scripts, and code for our primary analyses, are publicly accessible via our OSF project page (Link: https://osf.io/hdmbn/?view_only=0f65e6e7e85141bc836ef855cc2d7538). We report all

[8] Note that we also ran the multigroup SEM analyses for the other two outcome variables, the Bivalent gut reactions and actual feelings as well as the GEB.

measures, all manipulations, all data exclusions, and how we determined the sample size in this study. Deviations from the preregistration not reported in the main text are reported in Supplement 1. We conducted all analyses using R (version 4.2.2, R Core Team, 2021) and used the R package lavaan (Rosseel, 2012) for the SEM analyses.

Results

Preliminary analyses

Ensuring data quality. Following the recommendations of Greenwald et al. (2003) to ensure data quality when computing IAT effects based on the D score algorithm, we excluded participants who responded faster than 300 ms in 10% or more of the trials across all test blocks, and we excluded responses with latencies exceeding 10,000 ms. In addition, participants who did not complete the ST-IAT or the questionnaire were excluded.

Handling missing values, multivariate normal distribution, descriptive statistics, and testing measurement invariance. None of the observed variables had missing values. Because the observed variables were not multivariate normally distributed (Mardia's skewness = 182.38, $p < .001$; Mardia's kurtosis = 7.20, $p < .001$), we used the maximum likelihood mean-variance adjusted (MLMV) estimator. Descriptive statistics for all observed variables are provided in Supplement 1 on our OSF project page.[9] We assumed strong measurement invariance (MI), since only the chi square difference test between the strong MI and the strict MI model was highly significant (see Supplement 1 for the results of all model tests) and the strong MI model had a very good model fit, S-B $\chi^2_{\text{strong MI}}$ (21) = 20.82, $p = .470$; $\text{RMSEA}_{\text{strong MI}} = 0.00$; $\text{CFI}_{\text{strong MI}} = 1.00$; $\text{SRMR}_{\text{strong MI}} = 0.05$; $\text{AIC}_{\text{strong MI}} = 2026.5$; $\text{BIC}_{\text{strong MI}} = 2121.1$.[10]

[9] 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 preregistered SEM analyses.

[10] Following a suggestion of a reviewer we also used a Bayesian approach for testing MI and for testing our main analyses. We applied the information criterion leave-one-out cross-validation (LOO-CV; Vehtari et al., 2017) and used the absolute value of the LOO-CV difference, as indexed by the Expected Log Predictive Density (ELPD-LOO-CV), to compare models, as recommended by Vehtari et al. (2017). According to recommended guidelines, ELPD-

364 ***Main analyses***

365 **IAT test difficulty hypothesis.** As hypothesized, descriptively, the latent mean of the
366 UnivalentPos ST-IAT was closer to zero, $\hat{\mu}_{\text{UnivalentPos IAT}} = 0.25$, than the latent mean of the
367 Bivalent ST-IAT, $\hat{\mu}_{\text{Bivalent IAT}} = 0.45$ (see Table 1 for the latent means and their standard errors).
368 To test for significance, we added the group equality constraint to the strong MI model that the
369 latent means of the ST-IATs were equal. The resulting model (means model) fitted significantly
370 worse than the strong MI model (see the fit indices in Table 2), indicating that the latent means of
371 the ST-IATs differed significantly, which was also supported by a Wald tests, $W(1) = 17.61$, $p <$
372 .001.

373 In addition, Bayesian SEM model testing also suggests that the means model (ELPD-
374 LOO = -1024.2, *SE* = 30.5) makes worse predictions than the strong MI model (ELPD-LOO = -
375 1017.7, *SE* = 30.5), as the means model had a smaller ELPD-LOO and as the absolute ELPD-
376 LOO difference was above the threshold of four and exceeded its standard error (ΔELPD-LOO =
377 -6.6, *SE* = 3.7).

378 **IAT true-score variance hypothesis.** As hypothesized, descriptively, the true-score
379 variance of the UnivalentPos ST-IAT was higher, $\hat{\sigma}_{\text{UnivalentPos IAT}}^2 = 0.08$, than of the Bivalent ST-
380 IAT, $\hat{\sigma}_{\text{Bivalent IAT}}^2 = 0.05$ (see Table 1 for the true-score variances and their standard errors). To
381 test for significant differences in the true-score variances we added the group equality constraint
382 to the strong MI model that the true-score variances of the ST-IATs were equal. Contrary to our
383 expectation, however, the resulting model (variances model) did not fit significantly worse than

---

LOO differences below four indicate negligible differences between models, whereas differences above four can be considered evidence for differences between models if they exceed their standard error (Sivula, Magnusson, & Vehtari, 2020). In our MI analyses, the observed ELPD differences were all well below the threshold of four, indicating that the models did not differ meaningfully (all ΔELPDs <= –1.3, all *SEs* <= 1.5). As such, according to the Bayesian approach we could assume strict MI, however, since the frequentist approach suggests the strict MI model to fit significantly worse than the strong MI model, we decided to be conservative and only assumed strong MI for the following analyses to test our hypotheses. This ensures that we do not impose additional assumptions associated with strict MI that may not be warranted.

the strong MI model (see the fit indices in Table 2), indicating that the true-score variances of the ST-IATs were not significantly different, which was also supported by an individual Wald test, $W(1) = 2.81$, $p = .094$.

In addition, Bayesian SEM model testing also suggests that the variances model (ELPD-LOO = -1017.9, *SE* = 30.6) and the strong MI model (ELPD-LOO = -1017.6, *SE* = 30.4) make the same predictions, as the absolute ELPD-LOO difference was below the threshold of four and did not exceed its standard error (ΔELPD-LOO = -0.3, *SE* = 4.5).

IAT predictive power hypothesis. In contrast to our hypothesis the latent correlation of the UnivalentPos ST-IAT with the latent direct attitude variable was descriptively not only slightly smaller, $\hat{r}_{\text{UnivalentPos IAT}} = -.02$, than the corresponding latent correlation of the Bivalent ST-IAT, $\hat{r}_{\text{Bivalent IAT}} = .12$, but even approached zero (see Table 1 for the latent correlations and R^2). Although the difference in correlations points in the wrong direction and is very small, we have tested it for significance for the sake of completeness. Accordingly, we added the group equality constraint to the strong MI model that the latent covariances of the ST-IATs were equal. The resulting model (covariances model) did not fit significantly worse than the strong MI model (see the fit indices in Table 2), indicating that the latent covariances between the ST-IATs and the direct attitude measures were not significantly different from each other. Since neither the latent variances nor the latent covariances of the ST-IATs differed, it can be assumed that the latent correlations did not differ either, which was further supported by an individual Wald test, $W(1) = 0.54$, $p = .460$.

In addition, Bayesian SEM model testing also suggests that the covariances model (ELPD-LOO = -1016.9, *SE* = 30.5) and the strong MI model (ELPD-LOO = -1017.6, *SE* = 30.4) make the same predictions, as the absolute ELPD-LOO difference was below the threshold of four and did not exceed its standard error (ΔELPD-LOO = -0.6, *SE* = 0.7).

The pattern of results for the additional outcome variables was the same: For both outcome variables, the UnivalentPos ST-IAT showed weaker latent correlations (Bivalent gut reactions and actual feelings: $\hat{r}_{\text{UnivalentPos IAT}} = -.06$; GEB: $\hat{r}_{\text{UnivalentPos IAT}} = -.03$) than the Bivalent ST-IAT (Bivalent gut reactions and actual feelings: $\hat{r}_{\text{Bivalent IAT}} = .12$; GEB: $\hat{r}_{\text{Bivalent IAT}} = .07$). The latent correlations of the IATs with the respective outcome variables did not differ significantly in the frequentist analyses, nor was there any evidence of relevant differences in the Bayesian analyses (see Supplement 1 for a detailed description of the results).

Table 1

Latent Means, Latent True-Score Variances, Latent Correlations, R^2, and Reliabilities of the Latent IAT Variable for the Two Groups in the Strong Invariance Model (Experiment 1)

Group	L mean (*SE*)	L variance (*SE*)	L correlation (*CI*)	R^2	Reliability
Bivalent ST-IAT	0.45 (.03)	0.05 (.01)	.12 (-.16, .35)	0.014	.54
UnivalentPos ST-IAT	0.25 (.04)	0.08 (.02)	-.02 (-.31, .33)	0.000	.65

Note. L = latent; CI = bootstrap-bias-corrected confidence intervals; Bivalent ST-IAT = environmental protection single target implicit association test with Bivalent attribute categories; UnivalentPos ST-IAT = environmental protection single target implicit association test with positive Univalent attribute categories.

Table 2

Model Fit of the Different Models to Test the Overall Manipulation Hypotheses (Experiment 1)

Model	S-B χ^2 (df)	*p*	RMSEA	CFI	SRMR	AIC	BIC	$\Delta\chi^2$	*p*
Strong MI	20.82 (21)	.470	0.00	1.00	0.05	2026.5	2121.1		
Means	35.85 (23)	.043	0.08	0.89	0.09	2039.2	2127.2	16.70	<.001
Model	S-B χ^2 (df)	*p*	RMSEA	CFI	SRMR	AIC	BIC	$\Delta\chi^2$	*p*
Variances	24.63 (23)	.370	0.03	0.99	0.09	2029.3	2117.4	3.16	.206

Covariances	21.41 (22)	.495	0.00	1.00	0.05	2024.9	2116.3	0.49	.485

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.

Discussion

In Experiment 1, we manipulated the valence of the attribute categories of an ST-IAT with a target category that is universally evaluated as positive (environmental protection). As hypothesized the ST-IAT with univalent attribute categories was a) significantly closer to moderate test difficulty than the ST-IAT with bivalent attribute categories, but contrary to our hypotheses the UnivalentPos ST-IAT had b) neither significantly more true-score variance (although there was a non-significant tendency in that direction) nor c) more predictive power than the Bivalent ST-IAT.

One possible explanation for why we did not find evidence for the last two hypotheses is that undesirable recoding strategies in ST-IATs might have thwarted the effect of our manipulation. Since in the combined blocks of ST-IATs two categories are always paired and one category is always unpaired, participants can simplify the categorization task by categorizing the exemplar stimuli according to whether or not they belong to the unpaired category (e.g., if the categories environmental protection and good are assigned to one response key and the category bad is assigned the other response key, the task can be recoded by simply deciding whether the presented exemplar stimuli are bad or not bad; see, Rothermund & Wentura, 2010). Consequently, the exemplar stimuli would no longer be categorized according to their nominal

features (Rothermund & Wentura, 2004), which might not only thwart the effect of our manipulation but in general jeopardize the measurement of construct-relevant variance. On the one hand, this would explain why we found no significant differences between the true-score variances of the ST-IATs and, consequently between the I-E correlations, as the recoding strategies are in principle applicable to both ST-IATs, and on the other hand it would also explain why both correlations were close to zero, as measures that do not capture construct-relevant variance are unlikely to have predictive power.

Recoding strategies, however, might also explain why we found significant differences in test difficulty between the two ST-IATs. In the Bivalent ST-IAT the recoding strategies might be more easily applicable in one of the combined blocks than in the other (in one of the combined blocks recoding the task into bad vs. not bad is easy because the valence of the unpaired attribute category is opposite to that of the other two categories whereas this is not possible in the other combined block) whereas in the UnivalentPos ST-IAT recoding strategies are difficult to apply since in both combined blocks all three categories have qualitatively the same overall valence (which is positive).

In Experiment 2, we replaced ST-IATs with standard IATs. If the results of Experiment 1 are due to unwanted recoding strategies that are specific for ST-IATs, then the manipulation of the attribute categories in Experiment 2 should have the hypothesized effects on test difficulty, true score variance, and predictive power of the IATs.

Experiment 2

In Experiment 2, we developed three standard IATs. We stayed with the same content domain as in Experiment 1 and chose the target categories “environmental protection” and “environmental degradation” for all three IATs, again assuming that the former category would generally elicit positive evaluations, while the latter category would generally elicit negative

evaluations. The three IATs were a Bivalent IAT with the attribute categories negative vs. positive (from now on called Bivalent IAT), a Univalent IAT with the positive attribute categories positive vs. extremely positive (from now on called UnivalentPos IAT), and a Univalent IAT with the negative attribute categories extremely negative vs. negative (from now on called UnivalentNeg IAT). In addition to using a different IAT type, the subsequent inclusion of another IAT with negative Univalent attribute categories, and renaming the labels of the attribute categories, we also made smaller changes to the material in Experiment 2 compared to Experiment 1 (see the Measures section of Experiment 2).

Methods

Design and Procedure

We used a between design: the factor *attribute category* had three levels (Bivalent attribute categories vs. Univalent negative attribute categories vs. Univalent positive attribute categories), and the factor *block order* had two levels (compatible vs. incompatible block first). Participants were randomly assigned to the different conditions. They were then asked to provide their demographic data, complete the respective IAT, and answer a final questionnaire.

Sample

A total of $n = 180$ participants were recruited via Prolific. A prerequisite for participation in the study was that the participants' native language was German and that they were between 18 and 45 years old. Participants received money as compensation for their participation. None of the participants had to be eliminated for data quality reasons (see the Results section of Experiment 1 for a more detailed overview of the exclusion criteria). Consequently, the final sample consisted of 180 participants (67% male; 64% employed; 75% had a university degree in a subject other than psychology; mean age of $M = 31.08$ years [$SD = 6.3$]) who were almost evenly distributed across the three IATs (Bivalent IAT: $n = 60$; UnivalentNeg IAT: $n = 61$;

UnivalentPos IAT: n = 59). Although, according to a one-tailed a priori power analysis for z tests of two independent correlations with G*Power (alpha = 0.05, Power = 0.8, rho1 = .0, rho2 = .3) we should have recruited 133 participants per condition, we aimed for the given number of participants due to the available resources and preregistered that we would have continued data collection if at least one of the two Univalent IATs had a correlation that was at least .2 larger descriptively than the correlation of the Bivalent IAT (see the preregistration on our OSF page for more details). The expected effect sizes were based on Urban et al. (2024, study 3).[11]

Measures

Standard IATs. The three IATs only differed with regard to the attribute categories and the corresponding exemplar stimuli. The exemplars for the Bivalent IAT with the attribute categories negative vs. positive were four negative (i.e., lousy, nasty, awful, bad) and four positive (i.e., beautiful, great, good, friendly) adjectives; the exemplars for the UnivalentPos IAT with the attribute categories positive vs. extremely positive were four positive (i.e., fine, nice, pleasant, satisfied) and four extremely positive (i.e., ingenious, grand, outstanding, perfect) adjectives; the exemplars for the UnivalentNeg IAT with the attribute categories extremely negative vs. negative were four extremely negative (i.e., cruel, horrendous, catastrophic, terrible) and four negative (i.e., wrong, weak, hindering, low) adjectives. In contrast to Experiment 1 we chose different exemplars for the less extreme attribute categories of the Univalent IATs than for the same attribute categories of the Bivalent IAT. We made sure that the exemplars for the positive/negative attribute category of the Univalent IATs were slightly less positive/negative than the attribute categories themselves so that they are more distant in valence from the exemplars of the extreme attribute categories (in case of the Bivalent IAT the exemplars could

[11] A one-tailed sensitivity power analysis for z-tests of two independent correlations with G*Power (alpha = .05, power = .8, n_1 = 61, n_2 = 60) showed that we could detect an effect size of q = .46. Note that the sensitivity analysis is only a conservative estimate of the true sensitivity of our multigroup structural equation models.

also be slightly more positive or more negative than the attribute categories themselves). Apart from that, we selected the exemplars according to the criteria used in Experiment 1. The valence of the exemplars was assessed on the basis of another pretest in which a total of 56 participants rated 101 evaluative adjectives on an 11-point Bivalent scale with endpoints ranging from 1 (*extremely negative*) to 11 (*extremely positive*).[12]

The remaining description of the IATs applies to all three IATs. We used the standard IAT block structure by Greenwald et al. (2003). The target categories were environmental protection/environmental degradation. Compared to Experiment 1 we used pictures instead of words as target exemplars: four pictures for environmental protection (e.g., a picture depicting solar panels) and four pictures for environmental degradation (e.g., a picture depicting car fumes). Regarding the number of trials per block, we followed the suggestions of Greenwald et al. (2022) for four exemplar stimuli per category, resulting in 16, 16, 32, 48, 24, 32, and 48 trials for the respective seven blocks. Participants were instructed to respond to the exemplars by pressing the left (*D*) or the right (*L*) response key. Apart from these changes, the IAT procedure remained the same as in Experiment 1.

IAT effects were calculated based on the D score algorithm (Greenwald et al., 2003). Positive mean D scores indicate that participants evaluate environmental protection (environmental degradation) more positively (negatively) than environmental degradation (environmental protection).

Outcome variables. Similar to Study 1, we measured gut reactions and actual feelings via questionnaire items, only this time towards the two target categories. Accordingly the items were: (a) “Rate your gut reactions towards environmental protection”, (b) “Rate your actual feelings

[12] The adjectives "satisfied", "perfect" and "low" were not included in the pretest, but after extensive consideration of the available adjectives from the pretest, they appeared to be more suitable for selection.

towards environmental protection", (c) "Rate your gut reactions towards environmental degradation", and (d) "Rate your actual feelings towards environmental degradation". Items (a) and (b) were to be rated on a 10-point Univalent scale with endpoints ranging from 1 (*positive*) to 10 (*extremely positive*). Items (c) and (d) were to be rated on a 10-point Univalent scale with endpoints ranging from 1 (*extremely negative*) to 10 (*negative*). We then created difference scores between the ratings of the two target categories, once for gut reactions and once for actual feelings. Higher scores on the items indicate a more positive/negative evaluation of environmental protection/environmental degradation. On top of this central outcome variable, we collected three additional outcome variables. Firstly, Participants also had to rate all four items on a 10-point Bivalent scale with endpoints ranging from 1 (*extremely negative*) to 10 (*extremely positive*). We then again created difference scores as already described. Secondly and thirdly, we collected two more attitude measures, the general environmental attitude scale (Preisendörfer, 1999) and an ordering task in which participants were asked to rate environmental protection in terms of its valence relative to other positive words (i.e., friendship, justice, love, leisure, honesty, freedom; ratings of 7 indicate that participants rated environmental protection the most positively of all words, and ratings of 1 indicate that participants rated environmental protection the least positively).

Data analysis

Since the experimental design and hypotheses were very similar to those of Experiment 1, we again used multigroup SEM analyses to test our hypotheses, the main difference being that this time we had three IATs as experimental groups instead of two IATs. Accordingly, the model fitted in all groups was also very similar to that of Experiment 1. It consisted of two correlated latent variables. The latent D score variable, was again measured via three indicators, that is, three D scores that were created via parcels at the trial level. The latent direct attitude variable,

was also again measured via two indicators, but this time via the difference scores based on the Univalent gut reactions and the difference scores based on the Univalent actual feelings (for a conceptual representation of the model see Figure 1).[13] We controlled for potential block order effects in the same way as in Experiment 1.

Transparency and Openness

All points mentioned under the Transparency and Openness section of Experiment 1 also apply to Experiment 2.

Results

Preliminary analyses

Ensuring data quality. We used the same criteria to ensure data quality as in Experiment 1 (see the corresponding section of Experiment 1 for a detailed description).

Handling missing values, multivariate normal distribution, descriptive statistics, and testing measurement invariance. The observed variables had no missing values. Because the observed variables were not multivariate normally distributed (Mardia's skewness = 106.85, $p <$ .001; Mardia's kurtosis = 3.21, $p = .001$), we used the MLMV estimator. Descriptive statistics for all observed variables can be found in Supplement 2 on our OSF project page. As in Experiment 1, we assumed strong MI, since only the chi square difference test between the strong MI and the strict MI model was significant and the strong MI model had a very good fit, S-B $\chi^2_{\text{strong MI}}$ (36) = 41.75, $p = .235$; $\text{RMSEA}_{\text{strong MI}} = 0.05$; $\text{CFI}_{\text{strong MI}} = 0.98$; $\text{SRMR}_{\text{strong MI}} = 0.05$; $\text{AIC}_{\text{strong MI}}$ = 1979.4; $\text{BIC}_{\text{strong MI}} = 2103.9$ (see Supplement 2 for a more detailed description of the MI analysis).

Main analyses

[13] Note that we also ran the multigroup SEM analyses for the other three outcome variables, the Bivalent gut reactions and actual feelings, the general environmental attitude scale as well as the ordering task

IAT test difficulty hypothesis. As hypothesized, the latent means of the UnivalentNeg and the UnivalentPos IAT were descriptively closer to zero, $\hat{\mu}_{\text{UnivalentNeg IAT}} = 0.36$ and $\hat{\mu}_{\text{UnivalentPos IAT}} = 0.06$, respectively, than the latent mean of the Bivalent IAT, $\hat{\mu}_{\text{Bivalent IAT}} = 0.88$ (see Table 3 for the latent means and their standard errors). To test whether the latent means differed significantly we introduced the group equality constraint that the latent means of the IATs were equal into the strong MI model. The resulting model (means model) fitted significantly worse than the strong MI model (see the fit indices in Table 4). Consequently, the latent means of the IATs differed significantly overall. To test each latent mean difference, Wald tests were used, which showed that the Bivalent IAT differed significantly from the UnivalentNeg IAT, $W(1) = 121.07$, $p < .001$, as well as from the UnivalentPos IAT, $W(1) = 261.96$, $p < .001$, but interestingly also showed that the UnivalentNeg IAT differed significantly from the UnivalentPos IAT, $W(1) = 27.91$, $p < .001$.

In addition, Bayesian SEM model testing also suggests that the means model (ELPD-LOO = -1066, *SE* = 27.1) makes worse predictions than the strong MI model (ELPD-LOO = -996.5, *SE* = 28.6), as the means model had a smaller ELPD-LOO and as the absolute ELPD-LOO difference was way above the threshold of four and exceeded its standard error by far (ΔELPD-LOO = -69.5, *SE* = 9.8).

IAT true-score variance hypothesis. As hypothesized, the true-score variances of the UnivalentNeg and the UnivalentPos IAT were descriptively larger, $\hat{\sigma}_{\text{UnivalentNeg IAT}}^2 = 0.07$ and $\hat{\sigma}_{\text{UnivalentPos IAT}}^2 = 0.08$, respectively, than the true-score variance of the Bivalent IAT, $\hat{\sigma}_{\text{Bivalent IAT}}^2 = 0.04$ (see Table 3 for the true-score variances and their standard errors). To test for significant differences in the true-score variances we added a group equality constraint to the strong MI model, namely equal true-score variances in all groups. The resulting model (variances model) did not fit significantly worse than the strong MI model (see the fit indices in Table 4), indicating

that the true-score variances of the three IATs were not significantly different overall. This was also supported by individual Wald tests, which showed that none of the IATs differed significantly from each other in their true-score variance, as the Bivalent IAT differed neither significantly from the UnivalentNeg IAT, $W(1) = 3.14$, $p = .076$, nor from the UnivalentPos IAT, $W(1) = 3.49$, $p = .062$, and the Univalent IATs did not differ significantly from each other, $W(1) = 0.28$, $p = .596$.

In addition, Bayesian SEM model testing also suggests that the variances model (ELPD-LOO = -995.7, *SE* = 27.9) and the strong MI model (ELPD-LOO = -995.3, *SE* = 28.5) make the same predictions, as the absolute ELPD-LOO difference was below the threshold of four and did not exceed its standard error (ΔELPD-LOO = -0.4, *SE* = 3.7).

IAT predictive power hypothesis. In contrast to our hypothesis the latent correlation of the Bivalent IAT with the latent direct attitude variable was descriptively slightly larger, $\hat{r}_{\text{Bivalent IAT}} = .07$, than the latent correlations of the UnivalentNeg and the UnivalentPos IAT, $\hat{r}_{\text{UnivalentNeg IAT}} = -.02$ and $\hat{r}_{\text{UnivalentPos IAT}} = -.02$, respectively (see Table 3 for the latent correlations and R^2). Although the differences in correlations point in the wrong direction and are very small, we tested them for significance for the sake of completeness. We therefore added the group equality constraint to the strong MI model that the latent covariances of the IATs were equal. The resulting model (covariances model) did not fit significantly worse than the strong MI model (see the fit indices in Table 4). This result indicates that the latent covariances between the IATs and the direct attitude measures were not significantly different from each other. Since neither the latent variances nor the latent covariances of the IATs were significantly different, it can be assumed that the latent correlations did not differ either.

In addition, Bayesian SEM model testing also suggests that the covariances model (ELPD-LOO = -994.1, *SE* = 28.5) and the strong MI model (ELPD-LOO = -995.8, *SE* = 28.5)

make the same prediction, as the absolute ELPD-LOO difference was below the threshold of four (ΔELPD-LOO = -1.7, *SE* = 0.7).

The pattern of results for the additional outcome variables was the same: For all three outcome variables, the UnivalentNeg IAT and the UnivalentPos ST-IAT showed weaker latent correlations (Bivalent gut reactions and actual feelings: $\hat{r}_{\text{UnivalentNeg IAT}} = .08$, $\hat{r}_{\text{UnivalentPos IAT}} = -.01$; general environmental attitude scale: $\hat{r}_{\text{UnivalentNeg IAT}} = -.04$, $\hat{r}_{\text{UnivalentPos IAT}} = .03$; ordering task: $\hat{r}_{\text{UnivalentNeg IAT}} = -.05$, $\hat{r}_{\text{UnivalentPos IAT}} = .21$) than the Bivalent ST-IAT (Bivalent gut reactions and actual feelings: $\hat{r}_{\text{Bivalent IAT}} = .12$; general environmental attitude scale: $\hat{r}_{\text{Bivalent IAT}} = .25$; ordering task: $\hat{r}_{\text{Bivalent IAT}} = .35$). The latent correlations of the IATs with the respective outcome variables did not differ significantly in the frequentist analyses with the exception that the Bivalent IAT had a significantly larger latent correlation with the ordering task than the UnipoarNeg IAT, while there was no evidence for relevant differences in the Bayesian analyses (see Supplement 2 for a detailed description of the results).

Table 3

Latent Means, Latent True-Score Variances, Latent Correlations, R^2, and Reliabilities of the Latent IAT Variable for the Three Groups in the Strong Invariance Model (Experiment 2)

Group	L mean (*SE*)	L variance (*SE*)	L correlation (*CI*)	R^2	Reliability
Bivalent IAT	0.88 (.03)	0.04 (.01)	.07 (-.29, .43)	0.005	.78
UnivalentNeg IAT	0.36 (.04)	0.07 (.01)	-.02 (-.30, .25)	0.000	.72
UnivalentPos IAT	0.06 (.04)	0.08 (.02)	-.02 (-.26, .25)	0.000	.71

Note. L = latent; CI = bootstrap-bias-corrected confidence intervals; Bivalent IAT = environmental protection/environmental degradation implicit association test with Bivalent attribute categories; UnivalentNeg IAT = environmental protection/environmental degradation implicit association test with negative Univalent attribute categories; UnivalentPos IAT =

environmental protection/environmental degradation implicit association test with positive Univalent attribute categories.

Table 4

Model Fit of the Different Models to Test the Overall Manipulation Hypotheses (Experiment 2)

Model	S-B χ^2 (df)	*p*	RMSEA	CFI	SRMR	AIC	BIC	$\Delta\chi^2$	*p*
Strong MI	41.75 (36)	.235	0.05	0.98	0.05	1979.4	2103.9		
Means	194.32 (40)	<.001	0.25	0.39	0.72	2120.4	2232.1	148.35	<.001
Variances	50.34 (40)	.127	0.07	0.96	0.12	1981.3	2093.1	8.45	.076
Covariances	42.10 (38)	.298	0.04	0.98	0.05	1975.7	2093.8	0.32	.852

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.

Discussion

In Experiment 2, we manipulated the attribute categories of a standard IAT with the target categories environmental protection and environmental degradation, which can be assumed to be generally evaluated positively and negatively, respectively. As hypothesized, the IATs with Univalent attribute categories were a) significantly closer to moderate test difficulty than the IAT with Bivalent attribute categories, but contrary to our hypotheses, neither of the Univalent IATs had b) significantly larger true-score variance (although there was a non-significant descriptive trend in this direction) or c) more predictive power than the Bivalent IAT. As such, the pattern of the results was almost identical to that of Experiment 1.

Since we used traditional IATs in Experiment 2, which do not allow for the specific recoding strategies that might take place in ST-IATs (i.e., categorizing all exemplars in terms of whether they match or do not match the valence of the unpaired category), but nevertheless obtained almost identical results as in Experiment 1, it is unlikely that the results of Experiment 1 are caused solely by the use of these recoding strategies. Instead, the results of Experiment 2 further support the notion that manipulating the attribute categories does not have the intended effects. Admittedly, from a frequentist perspective it cannot be ruled out that the manipulation affects not only IAT test difficulty but also true-score variance, since the difference tests again reached marginal significance, similar to Experiment 1. From a Bayesian perspective, however, there was no evidence for differences in true-score variances, in either Experiment 1 or Experiment 2 (see the General Discussion for a further discussion of the matter). Crucially, though, regardless of whether the manipulation might also influence true-score variance, it certainly does not have the hypothesized positive influence on the predictive power of the IAT.

General Discussion

In two experiments, we attempted to positively influence the test difficulty, true-score variance, and predictive power of attitude IATs containing target categories that have a clear valence. For these clearly valenced targets, standard Bivalent IATs have low test difficulty, that is, very large mean IAT effects, linked to low true-score variance and low predictive power. Based on test theoretical considerations (Urban et al., 2024), we aimed to change the test difficulty towards moderately difficult IATs by manipulating the attribute categories from Bivalent to Univalent evaluative adjective pairs. In Experiment 1, we developed two ST-IATs with the target category environmental protection, one with Bivalent (bad vs. good) and one with positive Univalent (good vs. very good) attribute categories. In Experiment 2, we stayed with the same content domain and developed three standard IATs with the target categories environmental

protection/environmental degradation, one with Bivalent (negative vs. positive), one with positive Univalent (positive vs. extremely positive) and one with negative Univalent (extremely negative vs. negative) attribute categories.

In both experiments, in line with our hypotheses, the Univalent IATs showed test difficulties that were closer to moderate test difficulty than the test difficulties of the Bivalent IATs. However, both experiments also showed, contrary to our hypotheses, that the Univalent IATs had neither significantly more true-score variance nor significantly higher predictive power (assessed by I-E as well as I-C correlations) than the Bivalent IATs.

Admittedly, the possibility that the manipulation also influenced the true-score variance of attitude IATs, and thus the possibility of a type II error, cannot be completely ruled out, since all descriptive estimates point in the hypothesized direction and most difference tests reached marginal significance. However, also when using Bayesian analyses there was no evidence for relevant differences in true-score variances. In any case, the crucial point is that even if there were true-score variance differences the effects would be so small that it would require a sample size well beyond the usual sample size of IAT experiments to reliably detect it. Hogenboom et al. (2024) estimated the average sample size for IAT experiments to be between 65 and 81 participants based on work by Babchishin et al. (2013), Greenwald et al. (2009), and Oswald et al. (2013), which we already exceeded by far. Even more importantly, effects of the attribute categories on the test difficulty and, if any, on the true-score variance of the IATs have no downstream effects on the predictive power of the IATs, which is ultimately the relevant outcome. The question of whether or not there is an effect of attribute manipulation on true-score variance is therefore irrelevant to the practical conclusions and recommendations that we derive from our studies. The results clearly and unambiguously indicate that using Univalent instead of Bivalent attribute categories does not increase the predictive power of attitude IATs, which was

the main aim of the manipulation. On the contrary, the Bivalent IATs always showed (at least descriptively) higher correlations than the Univalent IATs, regardless of the experiment and the outcome variable. To strengthen our interpretation of the data we conducted a mini-meta-analyses across experiments and outcome variables. We used the correlations as effect size and ran a meta-regression to assess the moderating effect of IAT attribute category. We found the moderator to be highly significant, explaining 66.83% of the heterogeneity in effect sizes. Most importantly, the direction of the regression coefficient indicated that Bivalent IATs made better predictions than Univalent IATs (see Supplement 3 for a complete description of the meta-nalytical results). In other words, while on a study level the frequentist analyses as well as the Bayesian analyses suggest no differences between the correlations of the Univalent and the Bivalent IATs for the respective outcome variables (accept for the ordering task in the case of the frequentist analysis) across experiments and outcome variables we find that Univalent IATs are not only not better than Bivalent IATs, but are in fact significantly worse in making predictions. In the following, we discuss whether this central conclusion is admissible, by considering possible limitations of our experiments.

Limitations and alternative explanations of our results

It may seem as a limitation that we primarily focused on gut reactions and actual feelings as outcome variable to assess the predictive power of the IATs, since IATs were originally developed to predict behavior over and above such direct attitude measures. However, it can be argued that, on the one hand, most of the literature not only supports the notion that a relationship exists between IATs and direct attitude measures (e.g., Greenwald et al., 2009; Oswald et al., 2013), but that, on the other hand, it has also been shown that an increase in this relationship is associated with a higher correlation between IATs and behavioral measures (Greenwald et al., 2009). Thus, using direct attitude measures as a criterion to evaluate the validity of an IAT is a

viable strategy, which comes with the advantage that these measures can be constructed more easily, also with regard to conceptual correspondence with the respective IATs (see Hofmann et al., 2005, and Kurdi et al., 2019, on the importance of conceptual correspondence between IATs and outcome variables).

Moreover, to increase the generalizability of our results, in Experiment 1 we assessed people's environmental behavior using the GEB and in Experiment 2 we collected two additional direct attitude measures, the general environmental attitude scale, and the ordering task. When analyzing these outcome variables, the main results regarding our hypothesis remained the same, that is, the Univalent IATs did not have an increased predictive power compared to the Bivalent IATs (see Supplements 1 and 2 for the results). If anything, there was even some indication that the Bivalent IAT had more predictive power than the Univalent IATs at least with respect to the ordering task. Accordingly, although we focused on gut reactions and actual feelings as outcome variables for the reasons specified above, we collected and analyzed several outcome variables of different types and always found the same key result: Using Univalent instead of Bivalent attribute categories does not increase the predictive power of attitude IATs.

Another possible limitation of our experiments is that we investigated our hypotheses only within one particular content domain, that is, environmental protection/environmental degradation. The I-E or I-C correlations in this content domain may generally be too low for the effect of the manipulation to emerge. We may have inadvertently chosen an inappropriate content domain to test our hypotheses, which in principle does not allow for correlations between IATs and outcome variables, regardless of the attribute categories used, and another content domain which in principle does allow for correlations between IATs and outcome variables may have produced the hypothesized results. Although this alternative explanation cannot be ruled out, it seems unlikely to apply to our results, since it has already been shown that IATs can be

developed in this content domain which correlate significantly with direct attitude measures similar to those used in our main analyses (Urban et al., 2024, study 3). Furthermore, the Bivalent IAT correlated significantly with the ordering task, whereas the Univalent IATs did not correlate significantly with any of the outcome variables. It thus seems unlikely that a different content domain would change the main result that the manipulation of the attribute categories does not increase the predictive power of attitude IATs.

A final possible limitation and at the same time explanation for the results might be that, despite our efforts, the Univalent IATs did not meet the criterion that the exemplars should be easy to sort. To get to the bottom of this, we conducted an exploratory analysis in which we investigated the overall accuracy rate and the average response times of the IATs. These analyses revealed highly similar errors rates and average response times for the Uni- and Bivalent IATs (Exp. 1: mean accuracy = 92.4% [UnivalentPos ST-IAT] vs. 93.6% [Bivalent ST-IAT]; average RT = 894 ms [UnivalentPos ST-IAT] vs. 829 ms [Bivalent ST-IAT]; Exp. 2: mean accuracy = 92.7% [UnivalentNeg IAT] vs. 94.8% [UnivalentPos IAT] vs. 92.6% [Bivalent IAT]; average RT = 813 ms [UnivalentNeg IAT] vs. 754 ms [UnivalentPos IAT] vs. 758 ms [Bivalent IAT]). Thus, the Univalent IATs did not fail this criterion. Consequently, the lack of increase in predictive power for Unipolar IATs cannot be explained by them failing typical IAT criteria (Greenwald et al., 2022). Since unwanted recoding strategies specific to ST-IATs cannot explain this finding either, the most likely explanations are that the manipulation of the attribute categories a) did not increase the true-score variance strongly enough to produce positive downstream effects on correlations, b) increased the true-score variance sufficiently, however, this variance was unrelated to the outcome variables measured or c) elicited unknown unwanted processes that counteracted the effect of IAT test difficulty on true-score variance and correlation. It is unclear which of these explanations is most accurate. However, investigating the causes of why the

manipulation failed to increase predictive power was not the focus of our studies. The focus was on determining whether the use of Unipolar instead of Bipolar evaluative adjective pairs as attribute categories can increase the predictive power of attitude IATs with clearly positive or negative target categories. Based on our results, we can conclude that this is not the case.

A further argument in favor of the validity of our conclusion is that the same inference also has to be drawn in the context of questionnaire construction. In both Experiments, our exploratory analyses consistently showed that the Univalent items were closer to moderate difficulty and had more true-score variance than the Bivalent items, but they also consistently showed that the predictive power of the Univalent items (measured via the correlation with the IATs) was not increased (see our OSF page for the corresponding code of the analyses). The pattern of results for questionnaire construction was thus similar to that for IAT construction, with the difference that in this case the Univalent items clearly had more true-score variance than the Bivalent items. Most importantly, however, the finding that Univalent compared to Bivalent evaluative adjective pairs have no greater predictive power remained unchanged.[14]

Conclusion

We modified the task design of attitude IATs with target categories expected to elicit clearly positive or negative evaluations for most participants by changing the valence of the attribute categories from Bivalent to Univalent evaluative adjective pairs. Based on predictions derived from the test difficulty account (Urban et al., 2024), this manipulation should result in more moderate test difficulty, higher true-score variance, and higher I-E as well as I-C correlations, thus tackling the long-standing issue of the IATs' low predictive power. However, the results of our two experiments suggest that while Univalent IATs are indeed closer to

[14] Note that in contrast to the manipulation of the attribute valence of the IATs, we did not manipulate the scale valence of the items in a between design, but all participants answered both the Bivalent and Univalent items.

moderate test difficulty, this shift is not accompanied by an increase in true-score variance or predictive power. Although matching the attribute categories to the valence of the target categories seemed theoretically plausible and a promising approach to improve the measurement quality of the IAT, based on this evidence in conjunction with the detailed rebuttal of possible limitations of our experiments, we would advise against using Univalent evaluative adjective pairs as attribute categories in the future.

References

Axt, J., Buttrick, N., & Feng, R. Y. (2024). A Comparative Investigation of the Predictive Validity of Four Indirect Measures of Bias and Prejudice. *Personality & Social Psychology Bulletin*, *50*(6), 871–888. https://doi.org/10.1177/01461672221150229

Babchishin, K. M., Nunes, K. L., & Hermann, C. A. (2013). The Validity of Implicit Association Test (IAT) measures of sexual attraction to children: A meta-analysis. *Archives of Sexual Behavior*, *42*(3), 487–499. https://doi.org/10.1007/s10508-012-0022-8

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. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*(6), 977–997. https://doi.org/10.1002/ejsp.487

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

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

Gawronski, B., & Corneille, O. (2025). Unawareness of Attitudes, Their Environmental Causes, and Their Behavioral Effects. Annu. Rev. Psychol., *76*, 359–384. https://doi.org/10.1146/annurev-psych-051324-031037

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., & Lai, C. K. (2020). Implicit Social Cognition. *Annual Review of Psychology*, *71*, 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality & Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

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

Höhne, J. K., Krebs, D., & Kühnel, S.-M [Steffen-M] (2021). Measurement properties of completely and end labeled Univalent and Bivalent scales in Likert-type questions on income (in)equality. *Social Science Research*, *97*, Article 102544. https://doi.org/10.1016/j.ssresearch.2021.102544

Höhne, J. K., Krebs, D., & Kühnel, S.-M [Steffen-M.] (2022). Measuring Income (In)equality: Comparing Survey Questions With Univalent and Bivalent Scales in a Probability-Based

Online Panel. *Social Science Computer Review*, *40*(1), 108–123. https://doi.org/10.1177/0894439320902461

Kaiser, F. G., & Wilson, M. (2004). Goal-directed conservation behavior: The specific composition of a general performance. *Personality and Individual Differences*, *36*(7), 1531–1544. https://doi.org/10.1016/j.paid.2003.06.003

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, *91*(1), 16–32. https://doi.org/10.1037/0022-3514.91.1.16

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review*, *97*(2), 253–270. https://doi.org/10.1037//0033-295x.97.2.253

Kurdi, B., & Banaji, M. R. (2019). *Relationship between the Implicit Association Test and explicit measures of intergroup cognition: Data from the meta-analysis by Kurdi et al. (2018).* https://doi.org/10.31234/osf.io/vpcx8

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

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. (2025). Increasing the validity of implicit measures: New solutions for assessment, conceptualization, and action explanation. In J. A. Krosnick, T.

H. Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit bias and racism* (pp. 491–516). New York: Cambridge University Press.

Menold, N. (2021). Response Bias and Reliability in Verbal Agreement Rating Scales: Does Polarity and Verbalization of the Middle Category Matter? *Social Science Computer Review*, *39*(1), 130–147. https://doi.org/10.1177/0894439319847672

Menold, N., & Raykov, T. (2016). Can Reliability of Multiple Component Measuring Instruments Depend on Response Option Presentation Mode? *Educational and Psychological Measurement*, *76*(3), 454–469. https://doi.org/10.1177/0013164415593602

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press. https://www.researchgate.net/publication/242155666_The_Implicit_Association_Test_at_Age_7_A_Methodological_and_Conceptual_Review

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

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

Pratkanis, A. R. (1989). The cognitive representation of attitudes. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 71–98). Erlbaum.

Preisendörfer, P. (1999). *Umwelteinstellungen und Umweltverhalten in Deutschland: Empirische Befunde und Analysen auf der Grundlage der Bevölkerungsumfragen "Umweltbewußtsein in Deutschland 1991-1998"*. Leske + Budrich. https://doi.org/10.1007/978-3-663-11676-9

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416

Raccuia, L. (2016). Single-Target Implicit Association Tests (ST-IAT) Predict Voting Behavior of Decided and Undecided Voters in Swiss Referendums. *PLoS ONE*, *11*(10), e0163872. https://doi.org/10.1371/journal.pone.0163872

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

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139

Rothermund, K., & Wentura, D. (2010). It's Brief But Is It better? An Evaluation of the Brief Implicit Association Test. *Experimental Psychology*, *57*(3), 233–237. https://doi.org/10.1027/1618-3169/a000060

Urban, M., Rothermund, K., & Koch, T. (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*

938 *of Personality and Social Psychology*, *127*(1), 31–57.
939 https://doi.org/10.1037/pspa0000391
940 Wigboldus, D. H. J., Holland, R. W., & van Knippenberg, A. (2004). Single target implicit
941 associations. *Unpublished Manuscript*.