Behavior Research Methods
https://doi.org/10.3758/s13428-021-01592-8

# The good and the bad: Are some attribute words better than others in the Implicit Association Test?

Jordan R. Axt[1,2] · Tony Y. Feng[1] · Yoav Bar-Anan[3]

Accepted: 29 March 2021

## Abstract

The Implicit Association Test (IAT) is one of the most popular measures in psychological research. A lack of standardization across IATs has resulted in significant variability among stimuli used by researchers, including the positive and negative words used in evaluative IATs. Does the variability in attribute words in evaluative IATs produce unwanted variability in measurement quality across studies? The present work investigated the effect of evaluative stimuli across three studies using 13 IATs and over 60,000 participants. The 64 positive and negative words that we tested provided similar measurement quality. Further, measurement was satisfactory even in IATs that used only category labels as stimuli. These results suggest that common sense is probably a sufficient method for selection of evaluative stimuli in the IAT. For reasonable measurement quality, we recommend that researchers using evaluative IATs in English select words randomly from the set we tested in the present research.

**Keywords** Implicit Association Test · Reliability · Validity · Implicit attitudes

Research about implicit social cognition (Greenwald & Banaji, 1995) centers on unintentional, uncontrolled, cognitively efficient mental processes that influence behavior and judgment. While much of the interest in the field can be attributed to its theoretical novelty, another contributing factor is the generation of individual difference measures of implicit social cognition that can be easily adapted to many research contexts (see Gawronski & De Houwer, 2014 for a review). The most popular of these measures is the Implicit Association Test (IAT; Greenwald et al., 1998). The IAT is an indirect measure of cognitions (e.g., attitudes), inferred from performance in a task of categorizing two pairs of categories (e.g., flowers/insects, good/bad) with only two motor responses. A conservative estimate of the number of published studies using the IAT exceeds 4000 and would cover nearly all areas of psychology, such as research related to exercise (Forrest et al., 2016), eating disorders (Ahern et al., 2008), political judgment (Hawkins & Nosek, 2012), intergroup relations (Turner & Crisp, 2010), and consumer behavior (Gibson, 2008). In addition to its influence in the psychological literature, the IAT has also played a significant role in larger discussions related to prejudice and intergroup disparities, as over 30 million people have completed an IAT using the Project Implicit website (Ratliff & Smith, in press).

Aside from its adaptability to many research contexts, the popularity of the IAT may also be attributed to the large amount of work that has gone into exploring the measure's validity, ranging from issues like construct validity (Bar-Anan & Nosek, 2014; Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007), internal validity (Dasgupta et al., 2003; Nosek, Greenwald & Banaji, 2005), and predictive validity (Buttrick et al., 2020). Meta-analytic investigations have provided further evidence of the IAT's utility, such as in predicting variance in intergroup behavior (Kurdi et al., 2019). Notably, there is also a skeptical view on the validity of the IAT as a measure of individual differences in implicit social cognitions (e.g., Schimmack, 2019), and, more broadly, on the implicit construct and the leading theories proposed for the distinction between implicit and explicit constructs (Corneille & Hütter, 2020). While the value of the concept of implicit social cognition, and the IAT specifically, will remain a continued topic of discussion, researchers can draw from a wide range of existing studies and resources to justify their use of the IAT.

Within the vast literature about IAT methodology, little published work has studied the effect of individual stimuli

✉ Jordan R. Axt
jordan.axt@mcgill.ca

[1] Department of Psychology, McGill University, 2001 McGill College Ave, Montreal, Quebec H3A 1G1, Canada

[2] Project Implicit, Washington, DC, USA

[3] Tel Aviv University, Tel Aviv, Israel

Published online: 04 May 2021

on the measurement quality of the IAT. Perhaps as a result of this paucity in empirical research about stimulus selection, there is no consistency in the individual stimuli used in different IATs to represent the attributes (e.g., good and bad) or categories (e.g., Black people and White people) in the test. Indeed, a brief review of recent evaluative IATs (i.e., IATs with categories that reflect valence, such as good/bad) finds that the specific stimuli used to represent each attribute are either not listed (Chevance, Caudroit, Romain & Boiche, 2017; Hagiwara et al., 2016; Panzone et al., 2016; Conroy et al., 2010; Haider et al., 2015) or when listed, show little consistency. Indeed, as of this writing, the ten most recent papers using an evaluative IAT each used a unique combination of words to represent the positive and negative categories (see Table 1), with no individual word appearing in more than four of ten studies.

Such unaccounted variability across IAT stimuli has the potential to impact research outcomes. Most drastically, if specific stimuli or combinations of stimuli significantly improve or degrade measurement quality, then prior findings using the IAT may have limited generalizability. Similarly, if certain stimuli are particularly detrimental, then earlier studies that produced null results may have been misinterpreted; for example, if researchers found that the IAT failed to predict a behavioral outcome, then such a result could have been driven by the IAT's stimuli rather than an actual lack of relationship between the behavioral measure and the implicit construct purportedly measured by the IAT. Finally, a less severe but still consequential possibility is that even relatively small effects of measurement quality due to variance in stimuli could introduce unnecessary noise, thereby further minimizing the already small to moderate associations found between the IAT and many relevant outcomes (e.g., Buttrick et al., 2020; Forscher, Lai, et al., 2019). For these reasons, a more systematic investigation of the role of stimuli variability, with a focus on finding the most useful evaluative stimuli, would be valuable both for interpreting past research using the IAT and for future uses of the measure.

## Stimulus effects on the IAT

Prior work has produced conflicting evidence concerning the importance of individual stimuli in measurement quality in the IAT. In some studies, variance among stimuli had no impact on IAT performance. For instance, De Houwer (2001) found

**Table 1** Positively and negatively valenced stimuli for the 10 most recent evaluative IATs where such information is available

| Reference | Positive stimuli | Negative stimuli |
|---|---|---|
| Hughes, S., Mattavelli, S., Hussey, I. & De Houwer, J. (2020). The influence of extinction and counterconditioning procedures on operant evaluative conditioning and intersecting regularity effects. *Royal Society Open Science, 7*, 192085. | Delicious, tasty, nice, good, gorgeous, wonderful, yummy, pleasant. | Rotten, disgusting, nasty, horrid, sick, vomit, horrible, unpleasant |
| Dickter et al. (in press). Assessment of Sesame Street online autism resources: Impacts on parental implicit and explicit attitudes toward children with autism. *Autism*. | Marvelous, superb, pleasure, joyful, beautiful, glorious | Horrible, awful, tragic, agony, painful, terrible |
| Qiu and Zhang (2020). Make exercise easier: A brief intervention to influence implicit attitudes towards exercise and physical activity behavior. *Learning and Motivation*, *72*, 101660. | Sunshine, happy, carefree, charmed, relax, vitality, hopeful, passionate, attractive | Depressed, frustrated, painful, cumbersome, tedious, boredom, hopeless, annoying, lazy |
| Hall and Lee (2020). Marital attitudes and Implicit Associations Tests (IAT) among young adults. *Journal of Family Issues*. | Enjoyable, fun, pleasant, happy, satisfying | Unenjoyable, boring, unhappy, unpleasant, unsatisfying |
| Brailovskaia and Teichert (2020). "I like it" and "I need it": Relationship between implicit associations, flow, and addictive social media use. *Computers in Human Behavior*, *113*, 106509. | Brilliant, diamond, joy, truth, sunrise | Awkward, hate, failure, slum, stink |
| Scaife et al. (2020). To blame? The effects of moralized feedback on implicit racial bias. *Collabra: Psychology*. | Good, happy, joy, love, pleasure | Agony, bad, evil, hurt, nasty |
| Goddard et al. (2020). Unsafe bicyclist overtaking behavior in a simulated driving task: The role of implicit and explicit attitudes. *Accident Analysis & Prevention*, *144*, 105595. | Joyful, lovely, wonderful, beautiful, pleasant, happy | Painful, terrible, horrible, cruel, awful, agony |
| King and Auschaitrakul (2020). Affect-based nonconscious signaling: When do consumers prefer negative branding?. *Psychology & Marketing*. | Bright, champ, pride, star, talent | Brutal, cruel, python, shark, tornado |
| Piccirillo et al. (2020). Self-stigma toward nonsuicidal self-injury: An examination of implicit and explicit attitudes. *Suicide and Life-Threatening Behavior*. | Kind, considerate, caring, just, moral, generous, loving, trustworthy, honest, pure | Untrustworthy, evil, selfish, manipulative, dishonest, cruel, gross, deceptive, immodest, hate |
| Dai et al. (2020). Attention and memory biases for aggressive information in college students with fragile high self-esteem. *International Journal of Psychology*. | Success, honesty, health, honor, luck | Pain, failure, stupid, ugly, cruel |

that similar IAT effects emerged in a British-foreign evaluative IAT that used positive British names and negative foreign names (e.g., Princess Diana, Adolf Hitler) versus negative British names and positive foreign names (Margaret Thatcher, Albert Einstein), though the small sample size ($N = 28$ in a within-subjects design) suggests the study likely had low statistical power to detect any stimuli effects.

However, similar results were found in two studies totaling over 600 participants (Stieger et al., 2010). In a clever design, participants completed IATs or single-target IATs concerning associations between the category of the self and attributes of anxious versus calm using either (a) a predetermined set of words as stimuli, (b) stimuli that each participant chose from a larger pool of options as particularly representative of either anxiety of calmness, or (c) stimuli generated by each participant individually to represent anxiety and calmness. Results found that these variations on stimuli did not impact overall performance, internal reliability, or test-retest reliability. However, even a sample of 600 participants across three conditions provides relatively low statistical power (e.g., only 18% power to detect a small effect of Cohen's $q = .10$ for differences in test-retest reliability). Finally, a very high-powered investigation ($N > 40,000$) of the number of stimuli used to represent each category in IATs developed to measure Black-White attitudes, old-young attitudes and the gender-science stereotypes showed no decreases in measurement quality—assessed by overall effect sizes and IAT–self-report correlations—when using as few as two stimuli for either the target or attribute categories (Nosek et al., 2005), though this study did find that measurement quality decreased slightly when the IATs used a single stimulus that was identical to the category label.

These results contrast with other work that has found more substantive impacts of stimuli on the IAT. Several studies have illustrated the impact of valence when selecting stimuli to represent IAT categories (rather than the evaluative attributes). For example, compared to an IAT using general names associated with Black and White people (e.g., Tyrone, Josh) or an IAT using admired White people and disliked Black people as stimuli, positive associations towards White people were significantly reduced when stimuli were of admired Black people (e.g., Michael Jordan) and disliked White people (e.g., Timothy McVeigh; Govan & Williams, 2004; Mitchell et al., 2003).

Separate studies have found that the stimuli used to represent the attributes in IATs could impact task performance. For instance, among female participants, gender-power associations (i.e., the degree to which male versus female names were associated with the categories of potency versus weakness) were stronger when the stimuli representing weakness were more positive (e.g., delicate) compared to more negative (e.g., timid; Rudman et al., 2001). Other work manipulated the frequency of positive or negative stimuli within the IAT itself, as West German participants showed more ingroup favoritism on the IAT when positive stimuli occurred on 75% of attribute trials than when negative stimuli were shown on 75% of attribute trials (Bluemke & Fiedler, 2009). A final study found greater pro-White race IAT scores when using more recognizable White names as stimuli (Dasgupta et al., 2000).

Similarly, and most relevant to the present study, IAT performance can be influenced by the use of attribute stimuli that are already associated with the categories used. For example, ingroup preferences in IAT scores were higher among West German participants when the positive stimuli in the IAT drew from pre-existing positive stereotypes about West Germans (e.g., successful) and negative stimuli drew from pre-existing negative stereotypes about East Germans (e.g., xenophobic; Bluemke & Friese, 2006). Comparable results have been found in gender associations, as participants showed greater pro-female attitudes when the IAT used positive words that were stereotypically related to females (e.g., beautiful) compared to an IAT where positive words were stereotypically related to males (e.g., independent; Steffens & Plewe, 2001).

## The current work

Previous research suggests that variance in stimuli *can* impact IAT performance and measurement quality. However, stimuli effects on IAT performance have only been observed through relatively drastic manipulations, such as when category stimuli representing Black and White people in a race IAT were made to differ strongly in positivity or negativity (Govan & Williams, 2004; Mitchell et al., 2003), or when attribute stimuli were selected *because* of potential contaminating effects due to existing associations with the IAT categories (Bluemke & Friese, 2006; Steffens & Plewe, 2001). In these cases, researchers have sought to "stack the deck" in an attempt to test the boundary conditions wherein stimuli may become so problematic that quality of measurement on the IAT is impacted. What is less clear is the role of stimuli variance when researchers, like in the studies listed in Table 1, have the opposite goal—selecting stimuli with the hopes of minimizing measurement error. A notable lack of standardization across IATs has led to a large amount of diversity among stimulus sets, but the consequences of this diversity are currently unknown.

The present work explored this question with large samples and a wide range of topics. We focused specifically on the effect of variance in stimuli on measurement quality for evaluative IATs, which are the most popular form of IAT: a recent meta-analysis of interventions to change performance on measures of implicit associations found that 67% of studies that used the IAT assessed implicit evaluations (Forscher, Lai, et al., 2019). Our purpose was to test whether even when researchers choose attribute words in order to maximize measurement quality, some evaluative words produce better measurement quality than other words. If that is the case, our research could provide a list of words that are useful for maximizing measurement quality in evaluative IATs. Alternatively, if we find no consistent effects of the choice of words on measurement quality, future research could use the whole set of words tested in the present research.

We used the following indicators of measurement quality (see Bar-Anan & Nosek, 2014 for similar criteria): mean-level effects, known-groups differences, correlations with direct measures, and internal reliability. Next, we provide further justification for each of these criteria.

**Mean-level effects** A superior measure should be more sensitive to the assessed construct. Assuming a modal response tendency in the IAT that is interpreted as a preference for one group over another, such as for White versus Black people or straight versus gay people (Nosek et al., 2007), then measurement error can only weaken the ability to detect such preferences in implicit associations and therefore result in lower overall effect sizes. This assumption that measurement error will only weaken the overall effect size is common in prior research on the IAT's validity (see Bar-Anan & Nosek, 2014 and Nosek et al., 2005, for parallel reasoning). As a result, stimuli that produce larger overall IAT effects will indicate better measurement.

**Known-groups differences** Relatedly, a superior measure of a construct should be better able to detect variance between groups known to differ on that construct. Prior work suggests robust differences in the construct captured by indirect measures across the attitudinal domains included in our studies, such as race (Nosek et al., 2007), sexuality (Jost et al., 2004), and weight (Sabin et al., 2012). Again assuming that measurement error only reduces known effect sizes, the magnitude of such differences will be underestimated with greater measurement error. Therefore, measures that minimize error would increase the size of these known-groups differences, such as between gay and straight participants in sexuality IATs. Past work seeking to validate other IATs or indirect measures has used similar criteria (e.g., Axt et al., 2021; Nosek et al., 2014).

**Correlations with self-report** A better measure of a construct should maximize correlations with related measures due to reduced error, assuming the error between measures is uncorrelated (Nosek, Greenwald & Banaji, 2005). Given widespread evidence that the IAT and self-report measures assess distinct but related constructs (Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007), it is expected that the IATs used here will have a reliable (but not perfect) correlation with parallel measures of self-reported attitudes. Therefore, stronger correlations between the IAT and direct measures signal reductions in measurement error (see Axt, 2018 for another example of using correlations between the IAT and self-report variables as a way of assessing measurement error).

**Internal reliability** Higher internal reliability does not guarantee superior measurement of a construct, but all else equal, measures with greater internal reliability minimize error in assessment of the targeted construct (see Sriram & Greenwald, 2009 for a similar approach).

Across three studies, 13 IATs, and more than 60,000 participants, we examined the role of variance in stimuli on IAT measurement quality. Study 1 tested whether, across 64 different words, the presence or absence of any one stimulus was associated with greater or weaker measurement quality. In a more direct test, Study 2 compared the measurement quality of IATs that used the best performing words and worst performing words from Study 1. Finally, after Studies 1 and 2 found no noticeable effect of attribute words, Study 3 examined whether variability and relevance of attribute words to the attribute categories are of *any* importance to the measurement quality of the IAT by comparing a typical evaluative IAT with an IAT that used either only the attribute names as the evaluative stimuli or an IAT that used nonwords unrelated to the attributes as the evaluative stimuli.

# Study 1

In Study 1, participants completed IATs with randomly selected words from a larger pool of positive and negative words. We compared whether measurement quality varied as a function of the presence or absence of each specific word.

## Method

### Participants

We analyzed data from visitors completing evaluative IATs at the Project Implicit demonstration site (https://implicit.harvard.edu/implicit/takeatest.html). In these evaluative IATs, stimuli were randomly sampled from a pool of possible words. Specifically, participants ($N = 252{,}670$, $M_{\text{Age}} = 31.07$, $SD_{\text{Age}} = 13.34$, 64.3% female, 71.4% White) completing attitudinal IATs had the "Good" label populated with eight of a possible 32 positive words and the "Bad" label populated with eight of a possible 32 negative words (see online supplement at https://osf.io/fxe8q/?view_only=68f3dfeb3d6b4015a5487878c722219d for full list). These words were taken from stimuli used in prior research on evaluative IATs (Nosek, 2005) and had been originally chosen to be readily categorizable as positive or negative.

For the data analyzed in Study 1, we first selected eight topics involving attitudinal IATs (age, Arab-Muslim, disability, race, religion, sexuality, skin tone, weight). The Religion task randomly assigned participants to complete an IAT measuring implicit associations for either Christianity vs. Judaism, Christianity vs. Islam, or Judaism vs. Islam. In total, there were then ten IATs included in Study 1. For each IAT, we began downloading data starting in July 2018 and added data from prior months until we reached enough completed IAT sessions such that, for each word in the stimuli pool, there were a minimum of 4500 completed IATs with that word. The word that appeared in the smallest number of IATs (i.e., study sessions) in our data appeared in 4868 IATs and was absent from 15,275 IATs. This sample size then allowed for very high-powered tests, such as 95% power to detect a Cohen's $q$ effect of .06 when comparing correlations between the IAT with self-reported attitudes. In Studies 1–3, we excluded participants who responded faster than 300 ms in more than 10% of trials (Nosek, Greenwald & Banaji, 2005).

## Measures

See Table 2 for the labels used for all categories and attributes, as well as whether the category stimuli for each IAT consisted of images and/or words. The online supplement details category stimuli used in all studies. The procedure of the IATs and our scoring of the IAT scores followed those outlined in Greenwald et al. (2003). IATs were scored such that higher values indicated more positive associations with the dominant group (the group listed first in Table 2).

For each topic, self-reported attitudes were assessed by a single seven-point relative preference item (Axt, 2018); for example, self-reported weight attitudes were measured by an item ranging from −3 = "I strongly prefer fat people to thin people" to +3 = "I strongly prefer thin people to fat people" with a neutral midpoint of 0 = "I like fat people and thin people equally". See the online supplement for full text for each self-report preference item.

## Procedure

Participants completed the IAT and self-report attitude measure in a randomized order. Each topic also included a demographic questionnaire of varying length, and other self-report variables that were not included in analyses.

## Results

Internal reliability was calculated using the same method as Bar-Anan and Nosek (2014). Separate $D$ scores were calculated for (1) IAT blocks 3 and 4 (40 trials), (2) the first half of IAT blocks 6 and 7 (40 trials), and (3) the second half of IAT blocks 6 and 7 (40 trials), and these three scores were used to calculate a Cronbach's α (Cronbach, 1955).

**Mean-level effects** Table S1 in the online supplement presents the overall IAT effect size (Cohen's $d$) for when each word was or was not included in the IAT. Across all words, the largest average reduction across IATs comparing the presence versus absence of a word was $d_{diff} = -.02$ ("selfish"), and the largest average increase was $d_{diff} = .011$ ("scorn"). We also coded, for each IAT, whether the presence or absence of each word was associated with greater or weaker effect sizes. No word was associated with either stronger or weaker effect sizes across all tests; one word ("annoy") was associated with weaker effects in nine of ten IATs, and four words ("cheerful," "detest," "poison," "scorn") were associated with larger effects in eight of ten IATs.

**Table 2** Design of IATs used in Study 1

| Topic | Category labels | Category stimuli | Attribute labels |
|---|---|---|---|
| Age | Young people / Old people | Images | Good / Bad |
| Arab Muslim | Other people / Arab Muslims | Words (names) | Good / Bad |
| Disability | Abled persons / Disabled persons | Images | Good / Bad |
| Race | White people / Black people* *OR*<br>European Americans / African Americans | Images | Good / Bad |
| Religion 1 | Christianity / Judaism | Images and words | Good / Bad |
| Religion 2 | Christianity / Islam | Images and words | Good / Bad |
| Religion 3 | Judaism / Islam | Images and words | Good / Bad |
| Sexuality | Straight people / Gay people | Images and words | Good / Bad |
| Skin tone | Light-skinned people/ Dark-skinned people | Images | Good / Bad |
| Weight | Thin people / Fat people | Images | Good words / Bad words |

*Participants were randomly assigned to complete a race IAT with either set of labels. The self-report attitude item matched the labels used in the IAT

**Internal reliability** Table S2 in the online supplement presents the internal reliability coefficient $\alpha$ for the presence or absence of each word in each IAT. Across all words, the largest average reduction in reliability across IATs that did or did not include each word was $\alpha_{diff} = -.007$ ("disgust"), and the largest average increase was $\alpha_{diff} = .005$ ("hatred"). Across IATs, no word was associated with either stronger or weaker internal reliability across all tests; four words ("awful", "disgust", "humiliate", "magnificent") were associated with lower internal reliability in eight of ten IATs, and two words ("attractive", "beautiful") were associated with greater internal reliability in nine of ten IATs.

**Correlations with self-reported attitudes** Table S3 in the online supplement presents the correlation $r$ between the IAT $D$ score and self-reported attitudes when each word was or was not present in each IAT. Across all words, the largest average reduction in variance explained ($R^2$) comparing the presence versus absence was $R^2_{diff} = -.006$ ("delightful"), and the largest average increase was $R^2_{diff} = .006$ ("horrible"). Across IATs, no word was consistently associated with either stronger or weaker correlations with self-reported attitudes; three words ("awful", "delight", "delightful", "scorn") were associated with lower IAT–self-report correlations in eight of ten IATs, and one word ("lovely") was associated with stronger IAT–self-report correlations in eight of ten IATs.

**Consistency across metrics** Across the three metrics used to evaluate measurement quality, we investigated whether any single word was consistently associated with greater or weaker measurement quality. Specifically, we inspected whether any word was ranked in the top or bottom 25% of words for each criterion. These analyses found that none of the 64 words were ranked in the top or bottom 25% of words across the three metrics, suggesting that no word was consistently related with better or worse measurement quality among the criteria used in Study 1.

## Discussion

Using three criteria for measurement quality, no single word out of 64 was consistently associated with better or worse measurement across ten IATs. This is suggestive, though not conclusive, evidence that when stimuli are selected with the intent of avoiding any contaminating or problematic influences (e.g., pre-existing associations with the categories; Bluemke & Friese, 2006), variation in word choice is unlikely to have substantive effects on measurement quality. However, the design of Study 1 might not lend itself to a strong test of this hypothesis, as the stimuli representing "good" and "bad" were selected from a larger pool of words for each study session. As a result, any single word that could have reduced measurement quality may have been frequently used alongside other words that simultaneously improved measurement quality. In other words, the random nature of selecting stimuli for the IATs in Study 1 could have diluted the positive or negative measurement effects of any single word.

Study 2 sought to test the effect of attribute stimuli with a stronger manipulation. We assigned participants to complete IATs using the words most associated with greater or weaker measurement quality based on Study 1's results. If the null results of Study 1 were a consequence of the noise introduced through randomly selecting the other stimuli that was used with each target word, then combining the best and worst performing words should compound any possible effects and create a stronger test of the role of stimuli variation in IAT measurement.

## Study 2

### Method

#### Participants

Methods and analyses for Study 2 were preregistered at https://osf.io/wpu6n/?view_only=b469dcebf5be4679819efb92709b6b0b. We targeted a minimum sample size of 1000 eligible participants per IAT and stimulus set. Delays in removing the study led to a slight increase in sample size, though no analyses were completed until all data were collected.

A total of 16,783 eligible IATs were completed through the Project Implicit research pool from 8829 participants ($M_{\text{Age}} =$ 34.5, $SD = 14.6$, 72.2% White, 66.1% female). Participants could complete multiple study sessions. Only study sessions where a participant completed the same IAT a second time (or more) were excluded (11.6% of sessions). For each IAT, this sample size allowed for over 95% power to detect an effect of Cohen's $q = .15$ when comparing correlation strength and an effect of Cohen's $f = .085$ (Cohen's $d = .17$) when comparing the magnitude of known-groups differences. Data, materials, and analysis syntax for Studies 2 and 3 can be accessed at https://osf.io/ezj5t/?view_only=1781c05b04d54b829fd2eff67e0d429c.

#### Measures

**IATs** Participants were randomly assigned to complete IATs related to race, sexuality, age, weight, skin tone, and Arab Muslims using the same category labels as in Study 1. The one change was that the race IAT used only the category labels "White people" and "Black people". A politics IAT using the categories "Democrats" and Republicans" was also included, with category stimuli consisting of party logos and prominent members (e.g., Joe Biden, Ronald Reagan). In Study 2, all attribute labels were "Positive" and "Negative".

Within each topic, participants were randomly assigned to complete an IAT with "high-performing" or "low-performing" words based on the results of Study 1. To determine each stimulus set, all 64 Study 1 words were ranked on ability to (1) maximize overall effect sizes, (2) strengthen correlation with self-reported attitudes, and (3) heighten internal reliability. An average ranking was calculated for each word. The eight positive and eight negative words with the highest average ranking were assigned to the "high-performing" set, while the eight positive and eight negative words with the lowest average ranking were assigned to the "low-performing" set.

The high-performing words were: friend, smiling, adore, joyful, pleasure, friendship, happy, attractive, bothersome, poison, pain, nasty, dirty, hatred, rotten, horrific. The low-performing words were: cherish, glad, delightful, fabulous, fantastic, magnificent, terrific, triumph, hurtful, annoy, disgust, despise, horrible, awful, disaster, humiliate.

**Self-reported attitudes** Participants completed five self-reported evaluation items. For each topic, participants completed a single relative preference item as in Study 1 (e.g., −3 = I strongly prefer Black to White people, +3 = I strongly prefer White to Black people), two thermometer items ranging from 1 = strongly dislike to 7 = strongly like assessing liking of each group separately, and two slider items ranging from 1 = extremely negative to 100 = extremely positive assessing positivity towards each group separately. A composite measure of self-reported attitudes was calculated by creating separate difference scores from the liking and thermometer slider items then standardizing and averaging those two difference scores with the relative self-reported preference item (Axt, Bar-Anan & Vianello, 2020).

**Demographics** Upon registering for the research pool, participants reported a number of demographic details that we used for the known-groups analyses. Depending on the topic, additional demographic variables were added to allow for tests of known-groups differences: a seven-point measure of perceived skin tone (1 = very light, 7 = very dark), a five-point measure of identification as an Arab Muslim (1 = not at all, 5 = very much), an item about sexual orientation that allowed participants to identify as "heterosexual or straight" or "lesbian or gay", among other options, a seven-point measure of perceived weight status (1 = very underweight, 7 = very overweight), and a seven-point measure of strength of identification with Republicans versus Democrats (1= identify much more with Republicans, 7 = identify much more with Democrats). See the online supplement for full text of all demographic items.

### Procedure

Participants completed the IAT and self-report items in a random order. All added demographic items were completed immediately after the self-reported attitude items.

## Results

Given the large number of analyses included in Study 2, it was likely that several could reach statistical significance (i.e., $p < .05$) by chance. As a result, our preregistration outlined criteria that we believed would indicate substantive evidence that the stimuli manipulation impacted measurement quality. First, we would conclude that there are differences between the low-performing and high-performing stimuli if at least three of the seven tests found significant differences in the same direction when comparing (1) strength of correlations with self-reported evaluations, (2) degree of internal reliability, or (3) differences in the magnitude of known-groups differences. In addition, we would only conclude that there are significant differences between stimuli sets on measurement quality if (1) the average effect on correlations with self-report exceeded a small effect of Cohen's $q = .10$ (Cohen, 1988), (2) the average effect on internal reliability exceeded a difference of $\alpha = .05$, and (3) the average effect on known-groups differences exceeded a small effect of $d = .10$ (or $\eta_p^2 = .0025$ in an ANOVA).

**Correlations** All IATs were positively correlated with the parallel self-report attitude measure (all $r$s > .128, all $p$s < .001). Table 3 lists the sample size and strength of the correlation with self-report for each IAT and stimulus set, as well as the results of a Fisher's $Z$ test comparing the strength of correlations for each topic. Across the seven tests, there were no reliable differences between the high-performing and low-performing stimuli conditions.

**Internal reliability** Internal reliability was calculated using the same procedure as in Study 1. Table 4 lists the sample size and internal reliability for each IAT and stimulus set, as well as the results of a Feldt (1969) test comparing internal reliabilities for each topic. There were no reliable differences between the high-performing and low-performing stimuli conditions.[1]

[1] In exploratory analyses, we also estimated internal consistency using the correlation between the $D$ scores computed from blocks 3 and 6 with the $D$ score computed from blocks 4 and 7 (applying the Spearman–Brown correction for split-half correlations). This approach gives more weight to trials in IAT blocks 3 and 6, which is more similar to how the overall $D$ score is computed. None of the comparisons between conditions reached statistical significance, meaning overall conclusions were the same as when using $\alpha$ (see online supplement for full analyses).

**Table 3** Correlations with self-reported attitudes in Study 2

| Domain | High-performing set *r* | Low-performing set *r* | Fisher's *Z* |
|---|---|---|---|
| Age | .128 (*N*=1162) | .164 (*N* =1250) | *Z*=–.90, *p*= .368 |
| Arab-Muslim | .173 (*N* =1059) | .218 (*N* =1085) | *Z*=–1.08, *p*= .280 |
| Politics | .629 (*N* =1012) | .664 (*N* =1007) | *Z*=–1.35, *p*= .177 |
| Race | .301 (*N* =1283) | .244 (*N* =1282) | *Z*= 1.56, *p*= .119 |
| Sexuality | .371 (*N* =1258) | .368 (*N* =1229) | *Z*= .09, *p*= .928 |
| Skin tone | .185 (*N* =1268) | .169 (*N* =1205) | *Z*= .41, *p*= .682 |
| Weight | .239 (*N* =1216) | .243 (*N* =1232) | *Z*=–.10, *p*= .920 |

**Known-groups differences** Our preregistered classifications for known-groups differences compared (1) young (18–30) vs. old (50+) participants on the age IAT, (2) participants who identified at least "a little" as being an Arab Muslim versus those who "did not identify at all" on the Arab Muslim IAT, (3) participants who identified slightly, moderately, or much more with Democrats versus those who identified slightly, moderately, or much more with Republicans on the politics IAT, (4) Black versus White participants on the race IAT, (5) heterosexual or straight versus lesbian or gay participants on the sexuality IAT, (6) participants who identified as very light-, light-, or somewhat light-skinned versus those who identified as very dark, dark, or somewhat dark on the skin tone IAT, and (7) participants who identified as underweight (or neutral) versus overweight on the weight IAT. See the online supplement for descriptive statistics for each social group within each IAT and stimuli condition.

Table 5 presents the results of independent samples *t* tests of *D* scores between known groups on each IAT as well as the results of the interaction term in a 2 (Social group) × 2 (Stimulus set) ANOVA in each topic. Here, the interaction term estimated the likelihood that the known-groups differences was larger in one stimulus set than in the other. Five topics produced the expected known-groups differences within each stimulus set (e.g., differences in *D* scores between straight or heterosexual versus lesbian or gay participants). Within these five topics, only one interaction term was reliable; specifically, differences between White and Black participants' *D* scores were greater in the high-performing than low-performing stimuli condition.

Two IATs—those concerning age and Arab Muslim attitudes—failed to produce any group differences, making the results of the ANOVA interaction term difficult to interpret. In retrospect, these results are compatible with past work that found very weak relationships between participant age and indirectly measured age attitudes (e.g., Axt et al., 2014; Chopik & Giasson, 2017), and small average effects of more negative indirectly measured attitudes towards Arab Muslims among non-Arab Muslim participants (Buttrick et al., 2020).

## Discussion

As in Study 1, we found no consistent effect in Study 2 for the selection of positive and negative words on the measurement quality of the IAT, despite our attempt to use a stronger manipulation of word selection. The "high-performing" stimuli of Study 1 did not reliably produce stronger correlations with self-report, greater internal reliability, or larger differences between social groups known to differ in the measured attitudes. These high-powered null results provide more compelling evidence that variance in individual stimuli selected *without* the goal of introducing contaminating effects does not impact measurement quality.

**Table 4** IAT internal reliability in Study 2

| Domain | High-performing set α | Low-performing set α | Feldt's *W* |
|---|---|---|---|
| Age | .681 (*N* =1177) | .651 (*N*=1254) | *W*= 0.91, *p*= .059 |
| Arab-Muslim | .725 (*N*=1077) | .718 (*N* =1101) | *W*= 0.98, *p*= .339 |
| Politics | .869 (*N* =1024) | .862 (*N* =1025) | *W*= 0.95, *p*= .203 |
| Race | .723 (*N* =1305) | .709 (*N* =1294) | *W*= 0.95, *p*= .187 |
| Sexuality | .782 (*N*=1279) | .765 (*N* =1241) | *W*= 0.93, *p*= .091 |
| Skin tone | .723 (*N* =1285) | .704 (*N* =1217) | *W*= 0.94, *p*= .120 |
| Weight | .736 (*N* =1230) | .726 (*N* =1252) | *W*= 0.96, *p*= .257 |

**Table 5** Tests and comparisons of known-groups differences in Study 2

| ay | High-performing set comparison | Low-performing set comparison | ANOVA interaction |
|---|---|---|---|
| Age | $t(774)=–0.77, p=.442, d=–.06$ | $t(806)=1.68, p=.093, d=.13$ | $F(1, 1580)=2.95, p=.086, \eta_p^2=.002$ |
| Arab-Muslim | $t(1059)=1.01, p=.315, d=.09$ | $t(1087)=0.93, p=.351, d=.09$ | $F(1, 2146)=0.01, p=.939, \eta_p^2<.001$ |
| Politics | $t(840)=26.18, p<.001, d=2.00$ | $t(842)=27.11, p<.001, d=2.08$ | $F(1, 1682)=0.01, p=.945, \eta_p^2<.001$ |
| Race | $t(980)=8.22, p<.001, d=1.06$ | $t(1002)=5.16, p<.001, d=.59$ | $F(1, 1982)=7.07, p=.008, \eta_p^2=.004$ |
| Sexuality | $t(1114)=9.04, p<.001, d=1.20$ | $t(1076)=9.07, p<.001, d=1.15$ | $F(1, 2190)=0.001, p=.970, \eta_p^2<.001$ |
| Skin tone | $t(1045)=4.61, p<.001, d=.47$ | $t(990)=6.09, p<.001, d=.66$ | $F(1, 2035)=1.42, p=.234, \eta_p^2=.001$ |
| Weight | $t(1211)=3.29, p=.001, d=.19$ | $t(1226)=3.51, p<.001, d=.20$ | $F(1, 2437)=0.02, p=.892, \eta_p^2<.001$ |

It is possible that individual stimuli may matter for some IATs more than others; for instance, the race IAT showed greater known-groups differences when using the high-performing versus low-performing stimuli. Though additional data would be needed to confirm whether or not this finding reflects a false positive or a real effect that is just specific to the race IAT, the totality of evidence from Study 2 suggests that our stimulus set manipulation did not consistently impact measurement quality.

The results of Studies 1 and 2 suggest that the use of most stimuli is unlikely to severely impact IAT measurement quality, but a related question concerns the importance of variation in stimuli at all. Study 3 investigates this issue by comparing measurement quality among evaluative IATs that used multiple stimuli to represent the positive and negative attributes against two clearly inferior alternatives: (1) IATs that had no variation in attribute exemplars (i.e., the exemplars were only the attribute names) and (2) IATs that had attribute exemplars with no pre-existing association with the category (i.e., using totally unrelated letters to represent the attribute categories). This latter condition represents a particularly strong test regarding the importance of attribute stimuli, as it allows participants to easily engage in task recoding (Rothermund & Wentura, 2004). Specifically, participants can categorize these stimuli based on visual appearance, and any instructions to treat such stimuli as exemplars of the concepts of positive or negative could be intentionally disregarded, a process that would reduce the effect of associations between valence and the target attitude objects on performance in the IAT.

# Study 3

## Method

### Participants

Methods and analyses for Study 3 were preregistered at https://osf.io/48vz7/?view_only=c47e91f99b58497fb7f460b63509d436. We again targeted a minimum sample size of 1000 eligible participants per IAT and stimulus set. Delays in removing the study led to a slight increase in sample size, though no analyses were completed until all data were collected.

A total of 27,274 eligible IATs were completed through the Project Implicit research pool from 13,879 participants ($M_{\text{Age}=}$ 36.3, $SD$ = 14.2, 67.1% White, 67.2% female). Only study sessions where a participant completed the same IAT a second time (or more) were excluded (15.0% of sessions). This sample size allowed for a minimum of 95% power to detect an effect of Cohen's $q$ = .14 when comparing correlation strength and an effect of Cohen's $f$ = .077 ($d$ = .15) when comparing the magnitude of known-groups differences.

### Measures

**IATs** Participants were randomly assigned to complete IATs related to race, sexuality, politics, weight, food, and the environment. The race, sexuality, politics, and weight IATs had the same category stimuli and labels as in Study 2. The food IAT assessed associations concerning "Meat" and "Vegetables", with each category using seven color images of different meats or vegetables as stimuli (see online supplement). The environment IAT assessed attitudes towards the concepts "Urban" (items: busy, noise, city, building, skyscraper) and "Rural" (items: farm, country, fields, slow, quiet). In Study 3, the attribute labels for all IATs were "Good" and "Bad".

Participants completed IATs using one of three stimuli sets for the "Good" and "Bad" categories. In the *Words* condition, IATs used the same words as the high-performing condition in Study 2. In the *Good-Bad* condition, stimuli were only the words *Good, good, Bad, bad*. Finally, in the *Q-Z* condition, stimuli were only the letters *Q, q, Z, z*. In this condition, IAT instructions told participants to "pretend that the letter 'Q' means any good word" and to "pretend the letter 'Z' means any bad word."

**Self-reported attitudes** We measured self-reported attitudes with the same five-item format as in Study 2.

**Demographics** Demographics items related to known-groups differences in race, sexuality, politics, and weight were the

**Table 6** Correlations with self-reported attitudes in Study 3

| Domain | Q-Z $r$ | Good-Bad $r$ | Words $r$ | Good-Bad vs. Q-Z: Fisher's $Z$ | Words vs. Q-Z: Fisher's $Z$ | Words vs. Good-Bad: Fisher's $Z$ |
|---|---|---|---|---|---|---|
| Food | *.215 (N= 1445)* | **.402 (N=1542)** | .395 ($N$= 1528) | $Z$= 5.66, $p$ <.001 | $Z$= 5.43, $p$ <.001 | $Z$= −.23, $p$= .818 |
| Environment | *.182 (N= 1424)* | .396 ($N$= 1490) | **.503 (N=1520)** | $Z$= 6.33, $p$ <.001 | $Z$= 10.0, $p$ <.001 | $Z$= 3.68, $p$ <.001 |
| Politics | *.412 (N= 1281)* | .574 ($N$= 1319) | **.652 (N=1368)** | $Z$= 5.49, $p$ <.001 | $Z$= 8.75, $p$ <.001 | $Z$= 3.24, $p$= .001 |
| Race | *.126 (N= 1547)* | .251 ($N$= 1616) | **.304 (N=1690)** | $Z$= 3.65, $p$ <.001 | $Z$= 5.32, $p$ <.001 | $Z$= 1.65, $p$= .099 |
| Sexuality | *.194 (N= 1449)* | .348 ($N$= 1571) | **.378 (N=1575)** | $Z$= 4.75, $p$ <.001 | $Z$= 5.52, $p$ <.001 | $Z$= 0.97, $p$= .332 |
| Weight | *.071 (N= 1500)* | .128 ($N$= 1548) | **.242 (N=1535)** | $Z$= 1.59, $p$= .112 | $Z$= 4.84, $p$ <.001 | $Z$= 3.28, $p$= .001 |

*Note*: Within each domain, values in bold denote the strongest correlation, and values in italics denote the weakest correlation

same as in Study 2. In addition, participants who completed the food IAT responded to a single yes/no question about whether they identified as a vegetarian or vegan, and participants who completed the environment IAT responded to an item concerning the area in which they currently lived ("large city", "suburb of a large city", "medium-sized city", "suburb of a medium-sized city", "small town", "rural").

### Procedure

Participants completed the IAT and self-report items in a random order. All additional demographic items were completed immediately after the self-reported attitude items.

## Results

As in Study 2, our preregistration outlined several criteria that we believed would indicate substantive evidence that the manipulations to IAT stimuli consistently impacted measurement quality. First, we would conclude that any manipulation impacted measurement quality if at least three of the six tests found reliable differences in the same direction when comparing (1) strength of correlations with self-reported attitudes, (2) level of internal reliability, or (3) differences in the magnitude of known-groups differences. In addition, in order to conclude a substantive effect of our manipulation, results would need to show (1) the average effect on correlations with self-report exceeded a small effect of Cohen's $q$ = .10 (Cohen, 1988), (2) the average effect on internal reliability exceeded a difference of $\alpha$ = .05, and (3) the average effect on known-groups differences exceeded a small effect of $d$ = .10 (or $\eta_p^2$ = .0025).

**Correlations with self-reported attitudes** All IATs were positively correlated with the parallel self-reported attitude (all $r$s > .071, all $p$s < .006). Table 6 lists the sample size and strength of correlation with self-report for each IAT and stimulus manipulation, as well as the results of a Fisher's $Z$ test comparing the strength of correlations between all conditions.

Relative to the Q-Z condition, the Words condition produced reliably stronger correlations for all six topics, and the Good-Bad condition did so for five topics. The Words condition also showed stronger correlations with self-report than the Good-Bad condition for three topics. Meta-analyzing the Cohen's $q$ effect sizes of differences in correlations, the Q-Z condition showed weaker correlations with self-report than the Words condition (meta-analytic $q$ = .25, $p$ < .001) and the Good-Bad condition (meta-analytic $q$ = .17, $p$ < .001). Finally, while the Words condition showed evidence of on average stronger correlations with self-report than the Good-Bad condition (meta-analytic $q$ = .08, $p$ = .001), this effect size was lower than our preregistered threshold of $q$ = .10 for indicating substantive differences in correlations.

**Internal reliability** Table 7 presents the IAT internal reliability within each condition and topic, as well as the results of Feldt tests comparing level of internal reliability between all conditions. Notably, even IATs using only "Q" and "Z" as stimuli showed moderate levels of internal reliability (minimum $\alpha$ = .66, median $\alpha$ = .75), and even showed greater internal reliability than the Good-Bad condition for two of six topics. However, relative to the Good-Bad and Q-Z condition, the Words condition showed higher internal reliability for all six topics.

Following the method outlined by Feldt and Charter (2006), a weighted average across topics found that the Words condition ($\alpha$ = .81) had a higher internal reliability than the Good-Bad ($\alpha$ = .73) and Q-Z conditions ($\alpha$ = .73). This difference was higher than our preregistered criteria of a difference in $\alpha$ greater than .05 to indicate substantive effect of stimuli on internal reliability.[2]

[2] As in Study 2, we also estimated internal consistency in an exploratory analysis using the correlation between the $D$ scores computed from blocks 3 and 6 with the $D$ score computed from blocks 4 and 7. Overall conclusions were the same as when using $\alpha$; the Good-Bad and Q-Z conditions did not consistently differ in correlation strength across domain, but for each domain, the Words condition showed substantively greater internal consistency (Cohen's $q$ > .10) relative to both the Good-Bad and Q-Z conditions. See online supplement for full analyses.

**Table 7** IAT internal reliability in Study 3

| Domain | Q-Z α | Good-Bad α | Words α | Good-Bad vs. Q-Z: Feldt's *W* | Words vs. Q-Z: Feldt's *W* | Words vs. Good-Bad: Feldt's *W* |
|---|---|---|---|---|---|---|
| Food | .739 (*N*= 1456) | *.722 (N= 1556)* | **.816 (*N*=1541)** | *W*= 0.94, *p*= .111* | *W*= 0.71, *p*< .001 | *W*= 0.66, *p*< .001 |
| Environment | .774 (*N*= 1433) | *.742 (N= 1502)* | **.850 (*N*=1538)** | *W*= 0.88, *p*= .006* | *W*= 0.66, *p*< .001 | *W*= 0.58, *p*< .001 |
| Politics | *.784 (N= 1304)* | .823 (*N*= 1343) | **.884 (*N*=1392)** | *W*= 0.82, *p*< .001 | *W*= 0.54, *p*< .001 | *W*= 0.54, *p*< .001 |
| Race | .663 (*N*= 1559) | *.641 (N= 1636)* | **.754 (*N*=1703)** | *W*= 0.94, *p*= .104* | *W*= 0.73, *p*< .001 | *W*= 0.69, *p*< .001 |
| Sexuality | .751 (*N*= 1460) | *.727 (N= 1584)* | **.793 (*N*=1594)** | *W*= 0.91, *p*= .037* | *W*= 0.83, *p*< .001 | *W*= 0.76, *p*< .001 |
| Weight | .676 (*N*= 1511) | *.668 (N= 1563)* | **.746 (*N*=1554)** | *W*= 0.98, *p*= .317* | *W*= 0.78, *p*< .001 | *W*= 0.77, *p*< .001 |

*Note*: Within each domain, values in bold denote the highest reliability, and values in italics denote the lowest reliability effect

## Known-groups differences

For the topics of politics, race, sexuality, and weight, our classifications for known-groups differences were the same as in Study 2. In addition, we compared food IAT performance among participants who did versus did not self-identify as vegetarian or vegan and compared environment IAT performance among participants who reported living in a large city or suburb of a large city versus those who reported living in a small town or rural environment. Each IAT and stimulus condition produced the expected difference between social groups, with the one exception being the Q-Z condition failing to produce differences in weight IAT *D* scores between participants who identified as underweight versus overweight. See the online supplement for descriptive statistics for each social group within each IAT and stimuli condition.

Table 8 presents the results of independent samples *t* tests of *D* scores between known groups on each IAT, and the results of the interaction term in a 2 (Social group) × 3 (Stimulus set) ANOVA in each topic. As before, a reliable interaction term would suggest that one stimulus set caused a stronger effect of group membership on the IAT scores. For two topics, sexuality and weight, the size of known-group differences did not reliably differ across the three stimuli conditions. The remaining four topics found reliable social group by stimulus set interactions. Follow-up analyses revealed that, for race and politics, the Q-Z condition produced weaker group differences than the Good-Bad (Race: $p = .028$, $\eta_p^2 = .002$, Politics: $p < .001$, $\eta_p^2 = .013$) and the Words condition (Race: $p < .001$, $\eta_p^2 = .012$, Politics: $p < .001$, $\eta_p^2 = .047$). In turn, the Good-Bad condition produced weaker differences than the Words condition (Race: $p < .001$, $\eta_p^2 = .005$, Politics: $p < .001$, $\eta_p^2 = .013$). For the Rural-Urban IAT, the Words condition produced larger group differences than the Q-Z ($p < .001$, $\eta_p^2 = .008$) and Good-Bad condition ($p = .012$, $\eta_p^2 = .004$), while the Good-Bad and Q-Z condition did not reliably differ from each other ($p = .149$, $\eta_p^2 = .001$). Finally, for the meat-vegetables IAT, the Q-Z condition produced weaker group differences than the Words condition ($p = .004$, $\eta_p^2 = .003$), while the Q-Z condition and Good-Bad condition did not reliably differ ($p = .076$, $\eta_p^2 = .001$), and neither did the Good-Bad condition and Words condition ($p = .175$, $\eta_p^2 = .001$). In total, the Q-Z condition had weaker known-groups differences than the Words condition in four of six topics, and for three of six topics compared to the Good-Bad condition. The Words condition produced stronger known-groups differences for three of six topics compared to the Good-Bad condition.

Across all six topics, meta-analyses found that the Good-Bad condition was associated with greater differences between known groups than the Q-Z condition (meta-analytic $\eta_p^2 = .002$, $p = .004$). In addition, the Words condition produced greater group differences than either the Q-Z (meta-analytic $\eta_p^2 = .008$, $p = .003$) or the Good-Bad (meta-analytic $\eta_p^2 = .002$, $p = .005$) conditions. However, these effects should be considered quite small, as only the contrast between

**Table 8** Tests and comparisons of known-groups differences in Study 3

| Domain | Q-Z comparison | Good-Bad comparison | Words comparison | ANOVA interaction |
|---|---|---|---|---|
| Food | *t*(1439)=3.95,*p*<.001, *d*=*.36* | *t*(1537)=6.87,*p*<.001, *d*= .61 | *t*(1522)=7.12,*p*<.001, ***d*=.73** | *F*(2, 4498)=4.42, *p*= .012, $\eta_p^2$= .002 |
| Environment | *t*(853)=4.24,*p*<.001, d=.30 | *t*(872)=6.58,*p*<.001, *d*= .46 | *t*(897)=8.24,*p*<.001, ***d*=.57** | *F*(2, 2622)=7.98, *p*< .001, $\eta_p^2$= .006 |
| Politics | *t*(1085)=14.74,*p*<.001, d=1.04 | *t*(1089)=23.71,*p*<.001, *d*=1.61 | *t*(1125)=29.40,*p*<.001, ***d*=1.90** | *F*(2, 3299)=56.89, *p*< .001, $\eta_p^2$=.033 |
| Race | *t*(1193)=2.77, *p*=.006, d=.25 | *t*(1223)=6.49, *p*<.001, *d*= .55 | *t*(1302)=10.74, *p*<.001, ***d*=.90** | *F*(2, 3718)=16.78, *p*< .001, $\eta_p^2$=.009 |
| Sexuality | *t*(1273)=5.66,*p*<.001, d=.72 | *t*(1406)=6.31,*p*<.001, *d*= .81 | *t*(1389)=7.38,*p*<.001, ***d*=.95** | *F*(2, 4068)=1.62, *p*= .197, $\eta_p^2$= .001 |
| Weight | *t*(1492)=0.34, *p*=.735, *d*=*.02* | *t*(1539)=2.99, *p*=.003, ***d*=.15** | *t*(1533)=2.60, *p*=.009, *d*= .13 | *F*(2, 4564)=1.99, *p*= .137, $\eta_p^2$= .001 |

*Note*: Within each domain, values in bold denote the strongest effect, and values in italics denote the weakest effect the Words and Q-Z condition ($\eta_p^2 = .008$ is equivalent to $d = .18$) exceeded our preregistered criteria of $d = .10$ for substantive differences between manipulations. See the online supplement for full reporting of each follow-up ANOVA.

### Discussion

Compared to using stimuli with no variation or no meaningful association with the attribute labels, using varied stimuli improved measurement quality by increasing correlations with self-report, internal reliability, and known-groups differences. The advantage of varied stimuli versus simply using the attribute labels as stimuli was evident but much weaker, with effects that exceeded our prespecified criteria for evidence of a substantive effect on internal reliability but not in maximizing known-groups differences or increasing correlations with self-report. Finally, using the attribute labels as stimuli created superior measurement relative to using novel, unrelated stimuli on correlations with self-report, but the effects did not exceed the prespecified criteria for known-groups differences or internal reliability. The relatively strong performance of the Q-Z condition and its ability to produce a majority of the IAT effects found in the other conditions suggests that many participants followed the instructions to think of “Q” as positive and “Z” as negative, and participants did not naturally adopt a task-recoding strategy (treating the categories as “Q” and “Z” rather than positive and negative) when given the opportunity.

Taken together, Study 3’s results suggest that using individual stimuli improves IAT measurement quality but is not necessary to achieve satisfactory measurement. Indeed, using meaningless stimuli that had no pre-existing association with the attribute labels still produced satisfactory measurement—evident in outcomes like reliable correlations with self-reported attitudes, known-groups differences for five of six topics, and a median internal reliability of $\alpha = .75$. In short, given the only modest discrepancies between using individual words or unvaried stimuli that only reflected the attribute labels, it is unlikely that differences within the use of stimuli are a source of significant variation in IAT measurement.

## General discussion

Three studies investigated how measurement quality was affected by variation in the words chosen as stimuli to represent the positive and negative attributes in evaluative IATs. In Study 1, an archival analysis of ten evaluative IATs did not find a consistent effect of the presence or absence of any individual word on the overall IAT *D* scores, internal reliability, or correlations with self-reported attitudes. Similarly, in Study 2, the best performing set of words (numerically) from Study 1 did not produce better measurement quality than the worst performing set of words. In Study 3, using the attribute labels as the attribute stimuli was inferior to using a set of eight words for each evaluative category, although the decrement in the measurement quality of the IAT was not always substantial. Further, even a condition that used stimuli unrelated to the attribute labels produced acceptable levels of internal reliability, reliable correlations with self-reported attitudes, and expected differences in IAT performance based on participants’ demographics, ideology, or self-perceptions.

Taken together, results from these studies indicate that variation in the words selected as IAT stimuli does not appear to be a strong source of variation in IAT measurement. Based on previous research that found some effects of specific item stimuli on IAT performance (e.g., Bluemke & Friese, 2006; Govan & Williams, 2004; Rudman et al., 2001), we speculated that some evaluative words might be best suited for producing high measurement quality in the IAT. However, we found no evidence that this is the case. The present results are consistent with previous research that suggested that the category labels have a larger effect on measurement quality than the specific items categorized to these categories (Axt et al., 2021; Mitchell et al., 2003). Practically, our results reassure researchers looking to use the IAT that their results are unlikely to be overly influenced by specific evaluative stimuli in the IAT, so long as those stimuli are unambiguously associated with the relevant attributes and do not have clear confounds with the selected categories (e.g., Steffens & Plewe, 2001). Further reassurance that common sense is probably sufficient for a satisfactory choice of evaluative words comes from the fact that, in Study 3, even a rather unimaginative and restricted choice of a word stimulus that is identical to the attribute category labels did not result in a drastic decrease in the measure quality. For researchers who wish to use evaluative IATs in English, the present research then offers 64 equally suitable words (see online supplement for full list).

Our failure to find differences between evaluative IATs that used different attribute stimuli decreases the probability that this factor surreptitiously contributed to past findings, such as the modest correlation found between the IAT and measures of relevant behavior (e.g., Gawronski, 2019; Kurdi et al., 2019). Our results suggest that variation among the word stimuli chosen to represent attribute labels does not introduce a significant source of noise into the quality of IAT measurement and is unlikely to further suppress associations between the IAT and outcomes of interest. As a result, researchers seeking to better understand or maximize the association between the IAT and relevant criterion measures may look towards more structural components of study design, such as the degree of conceptual correspondence between the IAT and the measure of interest (Irving & Smith, 2020; Payne et al., 2008).

The present results do not suggest that the IAT is *insensitive* to the effect of stimuli. Past work clearly shows that the IAT can be influenced by manipulations to the stimuli used to represent specific categories or attributes. However, this work required substantial changes to such stimuli, often to the point of deliberately introducing confounds into the measure, such as by using images of White people that were widely detested and images of Black people that were (at the time) widely admired (Govan & Williams, 2004), or in using attribute words that were intentionally meant to have pre-existing associations with the categories used (e.g., using "beautiful" as a positive word when assessing implicit gender associations; Steffens & Plewe, 2001). Like these previous studies, our choice of stimuli in Study 3 influenced measurement quality, but that result required the drastic step of using stimuli that had no pre-existing association with the attributes. These prior studies are helpful in illustrating that introducing serious confounds into the stimuli can have serious effects on IAT performance, yet the present work more fully reveals the inverse finding—without major confounds in the selected stimuli, variation in stimuli have no substantial effects on IAT performance.

At the same time, our conclusions are limited to the fact that using common sense for selecting stimuli for the IAT is enough to achieve satisfactory measurement. It is unclear, however, whether specific informed selection methods, such as tailoring the positive or negative items to each attitude object, may produce even greater measurement quality. Prior studies on this topic, which used stimuli that had pre-existing associations with the categories used in the IAT, suffered from low statistical power and only included a single attitude domain (e.g., Steffens & Plewe, 2001). For instance, measurement quality may be improved on a race IAT that uses attribute items that refer to traits that are stereotypically associated with White or Black people. However, it is also possible that this approach could degrade measurement quality by changing the associations being measured, as completing an IAT with negative items that are stereotypically Black and positive items that are stereotypically White could temporarily strengthen anti-Black associations. This is a worthy direction for future research that may lead to advances on the validity of the IAT, though the present results suggest that such work is not required for finding satisfactory measurement. However, the small decrease (if any) in measurement quality that we found when using the category labels as the attribute words might suggest that improvement in the selection of attribute exemplars would not be easy to accomplish.

## Extending prior work on IAT measurement

In addition to the question of variability among IAT stimuli, these studies speak to other issues related to IAT measurement. For one, our results can shed light on prior discussions regarding the number of stimuli required per category in order to achieve satisfactory IAT measurement. A previous investigation (Nosek, Greenwald & Banaji, 2005) manipulated the number of stimuli in target and attribute categories across three IATs measuring either racial attitudes, age attitudes, or gender-science stereotypes. One version of the race IAT had six stimuli to represent each racial category (i.e., six images each of Black and White people) and a single stimulus to represent the attribute category (i.e., only using the category labels "Good" and "Bad), a design that is very similar to the Good-Bad condition in Study 3. Relative to versions of the race IAT that included more attribute stimuli, Nosek et al. (2005) found that using only a single stimulus did not produce large changes in the overall IAT effect or correlations with self-reported racial attitudes.

Though present results largely replicate these conclusions and extend them to a greater number of IATs, the larger sample sizes used here were also able to detect an effect of lower internal reliability when using only a single stimulus per attribute. These data more fully highlight that while including multiple stimuli per attribute category should improve measurement quality, it is not a requirement for achieving expected IAT effects. This finding might help to simplify the IAT (for example, when the participants have low language proficiency) with no serious cost in measurement quality. Further, assuming the present finding generalizes to other IAT categories, this might help researchers who struggle to find more than a couple of stimuli for the attribute categories (e.g., Socialism vs. Capitalism) or the target categories (e.g., word stimuli for two political parties of a similar ideology). As a result, this finding may expand the research topics to which the IAT can be effectively applied.

Notably, one shortcoming of this work is its inability to speak to differences in stimulus modality, such as in comparing IAT performance when using words versus images to represent a category. Prior work suggests that stimulus modality may influence IAT performance (e.g., Meissner & Rothermund, 2015). For instance, a single-category IAT produced stronger associations between tastiness words and desserts (versus vegetables) when food was represented as pictures versus words, and similar results occurred in an evaluative IAT measuring positive associations for desserts versus vegetables, though these effects were limited to participants who reported being on a diet (Carnevale et al., 2015).

One explanation for the impact of stimulus modality on performance concerns the level of representation. More specifically, images may induce more lower-level processing compared to words since images are more concrete representations of the category that activate less extraneous knowledge (e.g., Puce et al., 1996). Indeed, follow-up studies have manipulated level of representation using the same IAT modality; for example, Dutch participants showed more negative associations towards immigrants (versus natives) when IAT stimuli depicted groups of people (invoking higher-level representations) compared to when stimuli only depicted a single person at a time (Foroni & Bel-Bahar, 2010; see also Cooley & Payne, 2017). A similar process may explain why the IATs used in the present work were largely resistant to variation in individual (word) stimuli. The use of words may have facilitated higher-order processing of the stimuli and IAT attribute categories, and since many different words can unambiguously fit into the attributes used here (e.g., “positive” or “negative”), participants may have had little difficulty processing any of these words as representing each attribute. That is, using words as stimuli may allow participants to take a more expansive approach to the attribute labels that allows for a greater number of stimuli to fall under that attribute. Follow-up research on this topic may seek to test this account directly, such as by manipulating participants’ perceptions of what words best reflect a certain category as well as investigating whether similar effects emerge when using image stimuli.

The results of the Q-Z condition in Study 3 extend this notion of participants’ flexibility in the ability to categorize stimuli. Even when the stimuli had no pre-existing association with the attribute labels, participants were able to incorporate the stimuli into their representation of the attribute and produce IAT performance that had acceptable levels of internal reliability as well as expected patterns of known-groups differences and correlations with self-reported attitudes. These data are strong evidence that IAT performance is much more dependent on the attribute or category labels used to determine *how* stimuli are categorized than the specific stimuli used to represent the categories or attributes (e.g., Axt et al., 2021).

## Limitations and future directions

One clear limitation of this work mentioned previously is the somewhat narrow scope of our manipulations, as we did not examine the effects of stimulus variability using image stimuli or other instances of text stimuli, such as when words are used to represent categories (e.g., first names associated more with Black versus White people) or attributes other than positive or negative (e.g., words associated with danger and safety). Though prior work suggests that the choice between representing categories or attributes as images versus words may have an impact on IAT performance (Carnevale et al., 2015; Meissner & Rothermund, 2015), it is less clear whether variation *among* the images used in IATs substantively impacts measurement quality. Similarly, the effects found here are specific to evaluative IATs, and stimulus variation may play a role in IATs seeking to measure stereotypic associations, like that between gender and science versus arts (e.g., Zitelny et al., 2017). While the current results cannot rule out the possibility that stimulus variation is an important factor in IATs using images or those assessing stereotypic associations, we see no a priori reason to expect this lack of generalizability. Regardless, this line of research will only benefit by extending the question into IATs that use other stimulus modalities or measure other types of associations, as the flexibility shown by participants in adapting the Q-Z labels to an evaluative context may not necessarily extend to IATs seeking to measure more specific associations than a general positive vs. negative distinction.

Another possible concern might be that we tested only 64 words, but there are many more evaluative words. Indeed, it is theoretically possible that we missed some words that would perform better than the words we chose to test in the present research. However, this seems less likely when considering the relatively small effect, found in Study 3, of replacing the words with stimuli identical to the attribute category labels. The modest decrease in the measurement quality of the IAT in that condition suggests that even if we had increased the set of tested words in Studies 1 and 2, no substantial variability in measurement quality would have been found.

More generally, this investigation focused on only one indirect measure, the IAT. The question of how variation among stimuli impacts measurement quality could be extended into other forms of the IAT, such as the single-category IAT (Karpinski & Steinman, 2006) as well as other indirect measures, such as the Go No-Go Association Test (Nosek & Banaji, 2001) and evaluative priming (Fazio et al., 1986). Given similarities in performance across these tasks (Bar-Anan & Nosek, 2014), we would anticipate that other indirect measures would also not be overly influenced by variation within individual stimuli. However, it is still possible that variation among stimuli may impact some tasks more than others, especially given preliminary evidence that indirect measures of implicit associations may engage or rely on different psychological processes (Foroni & Semin, 2012).

## Conclusion

Despite its wide usage within psychological research, the stimuli used for the IAT show considerable variability across researchers. If measurement quality were overly influenced by individual stimuli, then many conclusions from IAT studies may not generalize to other forms of the test, and variation from stimuli could be a significant source of measurement error across IATs. The present work suggests this scenario to be an unlikely one, as measurement quality across 13 evaluative IATs was not impacted by variation among words used to represent the positive and negative attributes. In fact, there was evidence of only small and somewhat inconsistent decrease in measurement quality when using only a single stimulus that was redundant with the attribute label. These results highlight a greater need for researchers to focus on more conceptual and theoretical explanations for when the associations detected on an IAT develop, change over time, and do or do not predict behavior.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01592-8.

**Authors’ Note** All data and study materials are available at the project page on the Open Science Framework (https://osf.io/ezj5t/?view_only=1781c05b04d54b829fd2eff67e0d429c). All measures, manipulations, and exclusions in the studies are disclosed.

## Declarations

**Conflicting Interests** This research was partly supported by Project Implicit. J. R. Axt is Director of Data and Methodology for Project Implicit, Inc., a nonprofit organization with the mission to "develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender, or other factors." There are no other potential conflicts of interest with respect to authorship or the publication of this article.

## References

Ahern, A. L., Bennett, K. M., & Hetherington, M. M. (2008). Internalization of the ultra-thin ideal: positive implicit associations with underweight fashion models are associated with drive for thinness in young women. *Eating Disorders*, *16*, 294–307.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*, 896–906.

Axt, J.R., Conway, M.C., Westgate, E.C. & Buttrick, N.R. (2021). Implicit attitudes independently predict gender and transgender-related beliefs. *Personality and Social Psychology Bulletin, 47*, 257–274.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, *25*, 1804–1815.

Axt, J. R., Bar-Anan, Y., & Vianello, M. (2020). The relation between evaluation and racial categorization of emotional faces. *Social Psychological and Personality Science*, *11*, 196–206.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*, 668–688.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*, 1264–1272.

Bluemke, M., & Fiedler, K. (2009). Base rate effects on the IAT. *Consciousness and Cognition*, *18*, 1029–1038.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects?. *Journal of Experimental Social Psychology*, *42*, 163–176.

Brailovskaia, J., & Teichert, T. (2020). "I like it" and "I need it": Relationship between implicit associations, flow, and addictive social media use. *Computers in Human Behavior*, *113*, 106509.

Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of Implicit Association Tests. *Journal of Experimental Social Psychology*, *88*, 103941.

Carnevale, J. J., Fujita, K., Han, H. A., & Amit, E. (2015). Immersion versus transcendence: How pictures and words impact evaluative associations assessed by the Implicit Association Test. *Social Psychological and Personality Science*, *6*, 92–100.

Chevance, G., Caudroit, J., Romain, A. J., & Boiché, J. (2017). The adoption of physical activity and eating behaviors among persons with obesity and in the general population: the role of implicit attitudes within the Theory of Planned Behavior. *Psychology, Health & Medicine*, *22*, 319–324.

Chopik, W. J., & Giasson, H. L. (2017). Age differences in explicit and implicit age attitudes across the lifespan. *The Gerontologist*, *57*, S169–S177.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conroy, D. E., Hyde, A. L., Doerksen, S. E., & Ribeiro, N. F. (2010). Implicit attitudes and explicit motivation prospectively predict physical activity. *Annals of Behavioral Medicine*, *39*(2), 112–118.

Cooley, E., & Payne, B. K. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin*, *43*, 46–59.

Corneille, O., Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*, 212–232

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin, 52*, 177–193. https://doi.org/10.1037/h0044919.

Dai, J., Gao, H., Zhang, L., & Chen, H. (2020). Attention and memory biases for aggressive information in college students with fragile high self-esteem. *International Journal of Psychology*.

Dasgupta, N., Greenwald, A. G., & Banaji, M. R. (2003). The first ontological challenge to the IAT: Attitude or mere familiarity?. *Psychological Inquiry*, *14*, 238–243.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, *36*, 316–328.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*, 443–451.

Dickter, C. L., Burk, J. A., Anthony, L. G., Robertson, H. A., Verbalis, A., Seese, S., ... & Anthony, B. J. (in press). Assessment of Sesame Street online autism resources: Impacts on parental implicit and explicit attitudes toward children with autism. *Autism*.

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

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, *34*, 363–373.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*, 215–227.

Foroni, F., & Bel-Bahar, T. (2010). Picture-IAT versus Word-IAT: level of stimulus representation influences on the IAT. *European Journal of Social Psychology*, *40*, 321–337.

Foroni, F., & Semin, G. R. (2012). Not all implicit measures of attitudes are created equal: Evidence from an embodiment perspective. *Journal of Experimental Social Psychology*, *48*, 424–427.

Forrest, L. N., Smith, A. R., Fussner, L. M., Dodd, D. R., & Clerkin, E. M. (2016). Using implicit attitudes of exercise importance to predict explicit exercise dependence symptoms and exercise behaviors. *Psychology of Sport and Exercise*, *22*, 91–97.

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

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, *14*, 574–595.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (pp. 283–310). (2nd ed.). New York: Cambridge University Press.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*, 178–188.

Goddard, T., McDonald, A. D., Alambeigi, H., Kim, A. J., & Anderson, B. A. (2020). Unsafe bicyclist overtaking behavior in a simulated driving task: The role of implicit and explicit attitudes. *Accident Analysis & Prevention*, *144*, 105595.

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*, 357–365.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4–27.

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*, 1464–1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Haider, A. H., Schneider, E. B., Sriram, N., Scott, V. K., Swoboda, S. M., Zogg, C. K., ... & Freischlag, J. A. (2015). Unconscious race and class biases among registered nurses: vignette-based study using implicit association testing. *Journal of the American College of Surgeons*, *220*, 1077–1086.

Hagiwara, N., Dovidio, J. F., Eggly, S., & Penner, L. A. (2016). The effects of racial attitudes on affect and engagement in racially discordant medical interactions between non-Black physicians and Black patients. *Group Processes & Intergroup Relations*, *19*, 509–527.

Hall, S. S., & Lee, K. H. (2020). Marital attitudes and Implicit Associations Tests (IAT) among young adults. *Journal of Family Issues*.

Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin*, *38*, 1437–1452.

Hughes, S., Mattavelli, S., Hussey, I. & De Houwer, J. (2020). The influence of extinction and counterconditioning procedures on operant evaluative conditioning and intersecting regularity effects. *Royal Society Open Science, 7*, 192085.

Irving, L. H., & Smith, C. T. (2020). Measure what you are trying to predict: Applying the correspondence principle to the Implicit Association Test. *Journal of Experimental Social Psychology*, *86*, 103898.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, *25*, 881–919.

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*, 16–32.

King, D., & Auschaitrakul, S. (2020). Affect-based nonconscious signaling: When do consumers prefer negative branding?. *Psychology & Marketing*.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ... & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*, 569–586.

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the implicit association test. *Social Psychological and Personality Science*, *6*, 740–748.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*, 565–584.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*, 625–666.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*, 166–180.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88.

Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J., & Greenwald, A. G. (2014). Understanding and using the brief implicit association test: Recommended scoring procedures. *PloS one*, *9*, e110938.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test. *Experimental Psychology*, *54*, 14–29.

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

Panzone, L., Hilton, D., Sale, L., & Cohen, D. (2016). Socio-demographics, implicit attitudes, explicit attitudes, and sustainable consumption in supermarket shopping. *Journal of Economic Psychology*, *55*, 77–95.

Piccirillo, M. L., Burke, T. A., Moore-Berg, S. L., Alloy, L. B., & Heimberg, R. G. (2020). Self-stigma toward nonsuicidal self-injury: An examination of implicit and explicit attitudes. *Suicide and Life-Threatening Behavior*.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional MRI study. *Journal of Neuroscience*, *16*, 5205–5215.

Qiu, Y., & Zhang, G. (2020). Make exercise easier: A brief intervention to influence implicit attitudes towards exercise and physical activity behavior. *Learning and Motivation*, *72*, 101660.

Ratliff, K. & Smith, C.T. (in press). Lessons from two decades with Project Implicit. In J. Krosnick, T. Stark & A. Scott (Eds.), *A Handbook of Research on Implicit Bias and Racism*. APA Books.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*, 139–165.

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin*, *27*, 1164–1178.

Sabin, J. A., Marini, M., & Nosek, B. A. (2012). Implicit and explicit anti-fat bias among a large sample of medical doctors by BMI, race/ethnicity and gender. *PloS one*, *7*, e48448.

Scaife, R., Stafford, T., Bunge, A., & Holroyd, J. (2020). To blame? The effects of moralized feedback on implicit racial bias. *Collabra: Psychology*.

Schimmack, U. (2019). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691619863798.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*, 283–294.

Steffens, M. C., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für experimentelle Psychologie*, *48*, 123–134.

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*, 940–944.

Turner, R. N., & Crisp, R. J. (2010). Imagining intergroup contact reduces implicit prejudice. *British Journal of Social Psychology*, *49*, 129–142.

Zitelny, H., Shalom, M., & Bar-Anan, Y. (2017). What is the implicit gender-science stereotype? Exploring correlations between the gender-science IAT and self-report measures. *Social Psychological and Personality Science*, *8*, 719–735.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.