Behavior Research Methods (2024) 56:468–484
https://doi.org/10.3758/s13428-023-02060-1

Using facial expressions instead of response keys in the implicit association test

Yoav Bar-Anan[1] · **Ronen Hershman**[2,3]

Accepted: 5 January 2023 / Published online: 26 January 2023

Abstract
Previous research found that when people are instructed to smile toward liked objects and show negative facial expressions toward disliked objects, their facial response is faster and more intense than when they are required to smile toward disliked objects and express negative facial response toward liked objects. The present research tested a technologically innovative indirect evaluation measure that was based on that finding. Participants completed an implicit association test (IAT)—a common indirect measure of evaluation, responding with their emotional facial expressions, rather than by pressing response keys. In two web studies, using emotional facial expression detection through a webcam, we found that the Facial Response IAT (FR-IAT) is a reliable and valid measure of evaluations, comparable to the keyboard IAT. Because facial responses provide more information than key responses, pursuing future improvements of the FR-IAT's methodology, software, and data analysis is a promising direction for enhancing the quality of indirect evaluation measurement. The same methodology and technology may also enhance other indirect measures of evaluation and cognitive tests related to emotion and judgment.

Keywords Facial expressions · Implicit measures · Implicit association test · Implicit social cognition · Attitudes

It is often challenging to measure the variables that interest psychologists. Mental processes occur covertly, and verbal communication reveals only a fraction of those processes. Therefore, indirect measures that bypass communication are needed. In the present manuscript, we report on the development of a methodology that might enhance the indirect measurement of evaluation. Specifically, participants completed the most commonly used indirect measure of evaluation—the implicit association test (IAT; Greenwald et al., 1998)—by responding with facial expressions instead of pressing computer keys. To detect participants' facial expressions over the Internet, through a webcam, we used a publicly available JavaScript code and integrated it with a publicly available JavaScript library for running online studies. We then conducted studies to examine the validity of this new measure—the Facial Response Implicit Association Test (FR-IAT).

✉ Yoav Bar-Anan
baranan@tauex.tau.ac.il

[1] School of Psychological Sciences, Tel-Aviv University, 69978 Tel Aviv, Israel

[2] Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[3] Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Indirect evaluation measures

Evaluation is an expression of favorability, the most dominant meaning dimension (Osgood, 1962), and a central topic of research in psychology (Banaji & Heiphetz, 2010). For decades, the main tool for measuring evaluation was the questionnaire. People reported how much they like or dislike other people, social groups, products, or any other object. Because questionnaires are limited by people's understanding of the question and their ability and motivation to respond accurately, psychologists developed indirect measures of evaluation that do not require participants to report their evaluation of the target objects. One category of indirect measures, the so-called implicit measures (Brownstein et al., 2019; Gawronski & De Houwer, 2014; Smith & Ratliff, 2015), includes mostly tasks that require participants to categorize stimuli to categories based on the meaning of the stimuli. The ease of categorization in those tasks is sensitive to people's evaluation of the stimuli.

In the IAT, the indirect evaluation measure that is the focus of the present research, participants use two key responses to categorize items into four categories: two attribute categories (e.g., *Good* and *Bad*) and two target categories (e.g., *Dogs* and *Cats*). Participants are required to use the same key-response for items that belong to one attribute category (e.g., *Good*) and one target category (e.g., *Dogs*), and the other key-response for items of the other two categories (*Bad* and *Cats*). The measure is based on the idea that people would respond faster when the key assignment of the four categories is compatible (rather than incompatible) with their personal preference. For example, participants who prefer dogs over cats would respond faster when *Good* shares a key with *Dogs,* and *Bad* shares a key with *Cats* than in the other pairing condition (*Good*+*Cats*, *Bad*+*Dogs*).

Because performance in the IAT is thought to be sensitive to evaluation, numerous studies have utilized the IAT for the indirect measurement of evaluation (for reviews, see Greenwald et al., 2022; Kurdi et al., 2019). Because participants do not intend to evaluate the target objects, the IAT is often considered a measure of unintentional evaluation that might occur with little need for cognitive resources and might escape people's control (Friese et al., 2008; Hofmann et al., 2007, 2008; Wiers et al., 2009).

Although the IAT has been used in numerous studies, its validation as a useful tool for measuring (unintentional) evaluation is still ongoing (Vianello & Bar-Anan, 2021). Because the IAT is a method rather than a measure, each IAT developed to measure evaluations of specific objects must be validated separately, just as each questionnaire must be validated separately. Still, evidence regarding the validity of one IAT could inform the evaluation of other IATs because of their shared measurement method. There is good evidence that IATs are related to self-reported evaluation, and that this relation is stronger when it is reasonable to assume that self-reported evaluation of the target objects would be more similar to the unintentional evaluation of those objects (Bar-Anan & Nosek, 2014; Nosek, 2005). There is also good evidence that the IAT is a useful measure for predicting behavior, with incremental validity beyond self-reported evaluation, especially when the target objects are social groups that people are motivated to avoid evaluating negatively due to social norms (Greenwald et al., 2009). On the other hand, research has shown that factors other than evaluation sometimes influence the IAT (e.g., Brown-Iannuzzi et al., 2019; Rothermund & Wentura, 2004; Uhlmann et al., 2006), and that IATs sometimes suffer from weak validity, showing only weak relation with measures of variables that are supposed to be related to evaluation (Greenwald et al., 2009; Kurdi et al., 2019). In the context of the present research, these limitations are good justifications for research innovations that might improve measurement quality.

Emotional facial expressions

Humans tend to show facial expressions when they feel or want to convey certain emotions (Bavelas & Chovil, 1997; Parkinson, 2005). For example, most people smile to convey positive emotions (Messinger et al., 1999). Therefore, people's spontaneous emotional facial expressions when perceiving a stimulus may reveal their evaluation of the stimulus. Indeed, research that measured spontaneous emotional facial expressions when seeing Black people predicted anti-Black discrimination behaviors (Vanman et al., 2004, 2013). However, measuring spontaneous facial expressions has not become a common measure of evaluation, possibly due to technological barriers (the extant research mostly recorded facial electromyography), or for lack of strong and consistent evidence for its validity. After all, people might not always spontaneously display facial expression that match their evaluation of the stimuli presented to them (Fernandez-Dols et al., 1997; Russell et al., 2003).

For the purpose of indirect measurement of evaluation, it might be useful to move from a focus on detection of *spontaneous* emotional facial expressions to the measurement of *required* facial expression responses. When people are instructed to show a specific emotional facial expression upon seeing stimuli of a certain category, the emotional expression is faster and stronger (in the amplitude of the muscle contraction) when the emotion is compatible (rather than incompatible) with the normative valence of the category. Participants showed a faster and stronger facial response when they were instructed to smile toward flowers and frown toward snakes, than when instructed to smile toward snakes and frown toward flowers (Dimberg et al., 2002). Another study found initial evidence that this effect also occurs for subjectively liked and disliked objects. Participants categorized photos based on the spatial orientation of the image (tilted or upright) using smiling and frowning as the categorization responses. Half of the photos showed exercising and half showed sedentary office work. The participants who were faster to frown toward non-exercise photos, in comparison to their frowning speed toward exercise photos, tended to report a higher number of weekly exercise sessions (Brand & Ulrich, 2019). Based on this initial evidence, in the present research, we developed an IAT with emotional facial responses, and tested its viability as an indirect measure of evaluation.

Augmenting the IAT with emotional facial responding

We developed the Facial Response IAT (FR-IAT) by replacing the key-response associated with each attribute category with the emotional facial response that corresponds to that attribute. Table 1 shows the structure of the task. In the first block of the FR-IAT, participants had to smile when they saw positive words and to show a negative facial expression (expressing anger, sadness, or disgust) when they saw negative words. In Block 2, participants had to smile when seeing a photo of an item that belongs to one (randomly selected) target category and show a negative expression when seeing items of the other target category. In the original IAT, participants practice categorizing stimuli to the target categories in Block 1, and practice categorizing the attribute stimuli to the attribute categories in Block 2. In the FR-IAT, we reversed the order of these blocks, to reaffirm to participants that positive stimuli are categorized with a smile and negative stimuli are categorized with negative facial expressions. We wanted to increase the likelihood that performance in Block 2 would be related to participants' evaluation of the target categories. We expected them to respond more easily in Block 2, if their preferred target category required smiling and their less-preferred target category required a negative response, than if the opposite was required.

As in the original IAT, Blocks 3 and 4 of the FR-IAT were the combined task—a combination of Blocks 1 and 2: stimuli of all four categories appeared and participants had to respond to them with the same response required in Block 1 (smile toward positive words, respond negatively to negative words) and Block 2 (smile toward photos representing one target category, respond negatively toward photos representing the other target category). In the original IAT, performance in these blocks is thought to be sensitive to how strongly people associate the target category with the attribute category with which it shares a key response. In the FR-IAT, we thought that the effect of pairing the target categories with the attributes by a shared response might be augmented by the fact that the response (emotional facial expression) is strongly related to the same valence as the attribute category.

In Block 5, as in the original IAT, participants practiced reversing the responses required for categorizing each target category. Stimuli of the target category that required smiling in Block 2 required a negative response in Block 5, while stimuli of the target category that required the negative response in Block 2 required smiling in Block 5. As we explain later, we thought that a comparison between the performance in Block 2 and Block 5 would be sensitive to the participants' evaluation of the target categories.

Like in the original IAT, Blocks 6 and 7 were a combination of Blocks 1 and 5. Stimuli of all four categories appeared and participants had to respond to them with the same response required in Block 1 (smile to positive words, respond negatively to negative words) and Block 5 (respond negatively to stimuli of one target category and smile toward stimuli of the other target category).

The potential of the FR-IAT to improve measurement quality, in comparison to the original IAT, is based on two main features that characterize emotional facial responses but not keyboard responses: the pre-existing association between valence and facial expressions, and the possibility of measuring multiple features of the emotional facial responses, rather than only the response latency. The expectation that the pre-existing association between valence and facial expression would influence performance in the FR-IAT is compatible with findings from other tasks that detected evaluation by using relevant stimulus–response compatibility (De Houwer, 2003). Such tasks are designed to measure the compatibility between the to-be-categorized feature of the stimulus (e.g., whether it is a dog or a cat) and the response itself (e.g., whether it is a positive or a negative facial response). Previous research found such compatibility effects when the response was uttering the words *good* or

Table 1 The structure of the FR-IAT and the IAT

Task	FR-IAT		IAT	
Response	Negative response	Smile	Left key	Right key
Block				
1	Bad words	Good words	Cats	Dogs
2	Cats	Dogs	Bad words	Good words
3+4	Bad words, Cats	Good words, Dogs	Bad words, Cats	Good words, Dogs
5	Dogs	Cats	Dogs	Cats
6+7	Bad words, Dogs	Good words, Cats	Bad words, Dogs	Good words, Cats

The categories Cats and Dogs are an example from Study 1. The order of the pairing is typically counterbalanced between participants: half of the participants start with Cats sharing a response with Good words, and the other half start with Cats sharing a response with Bad words

bad (Voß et al., 2003), or performing approach or avoidance movements (e.g., Field et al., 2006). Arguably, the pre-existing association between facial responses and evaluation is stronger than the responses used in previous research, which might result with stronger compatibility effects and a higher sensitivity to evaluation.

The other potential advantage of the FR-IAT over the IAT and other indirect evaluation measures is that a facial response is richer in features than a keyboard response. The only data analyzed from the original IAT is the latency for correct responses. If a person is faster in pressing the correct key in Blocks 3 and 4 (e.g., when *Good words* and *Dogs* share one key and *Bad words* and *Cats* share the other key) in comparison with Blocks 6 and 7 (e.g., when *Good words* and *Cats* share one key and *Bad words* and *Dogs* share the other key), that difference is considered the person's indirectly measured preference (e.g., for dogs over cats). In the FR-IAT, the facial response in each trial can be tracked over time, from the onset of the stimulus until the end of the trial. Tracking response dynamics over time has already proven informative of cognitive processes that occur before the final response decision in tasks that utilized mouse movement responses (e.g., Smeding et al., 2016). For example, the mouse movement until the participants click the labels "Like" or "Dislike" to indicate their evaluation of a target object may reveal the level of ambivalence toward the evaluated objects (Schneider & Mattes, 2021). The response dynamics in facial response are not merely two-dimensional such as mouse movement (i.e., providing x and y values at any given time). Rather, facial expressions are each a combination of several distinct muscles, and each may be activated at any given time after the onset of the target (to-be-categorized) stimulus. For example, if the task rules require the participant to smile towards a stimulus that the participant dislikes, the participant might initially show a fleeting negative response before smiling. Further, facial responses differ not only by how quickly they are shown, but also by their development over time, by their (peak and average) strength, and by distinct visible qualities (e.g., a genuine smile versus a polite smile). Although the present research focused mostly on the initial validation of the FR-IAT, based on relatively simple features (latency and intensity), these potential advantages provide clear path for improving the informative value of the FR-IAT, in comparison with other indirect measures of evaluation.

The technology

Shortly after the invention of the IAT as a computerized task, a web version of that task was developed, and enabled collecting data from millions of people (Nosek et al., 2000; Ratliff & Smith, 2021). This massive trove of data has helped improve the methodology of the IAT (e.g., Axt et al., 2021; Greenwald et al., 2003), and the knowledge about its psychometric qualities (e.g., Bar-Anan & Nosek, 2014; Nosek et al., 2005). There are currently a few technological solutions for running the IAT over the Internet, some of them publicly and freely available (e.g., Carpenter et al., 2019). To provide similar advantages for the FR-IAT, we searched for freely available technological solution that would allow administering the FR-IAT online. We searched for a technology that provides immediate detection of facial expressions from a video stream captured by a webcam, to provide participants immediate feedback about their response. We preferred not to record videos, to ensure the participants' privacy and to avoid burdening the network and the data server with large records.

We used the openly available face-api.js library (https://justadudewhohacks.github.io/face-api.js/), which provides a facial expression recognition module, based on an algorithm trained on variety of images from publicly available datasets. The library is programmed in JavaScript, running on the participant's browser (i.e., client-side). The library detects, in real time, the face in the image captured by the webcam. The facial expression recognition module provides, for each of seven possible expressions (happy, sad, disgusted, afraid, surprised, angry, or neutral), an estimate of the probability that the target face currently displays that expression. The estimation is relative, such that the sum of all seven probabilities in each detection attempt is always 1. In other words, the software tries to detect the currently dominant emotional facial expression, and its output is the probability that each of the seven emotional facial expressions is currently the dominant expression. The library does not distinguish between different subtypes of facial expressions (e.g., genuine vs. polite smile) and the only indication it provides for the intensity of the emotional expression is the probability that it is the dominant expression, which is probably only partly correlated with expression intensity. These aspects are suboptimal, leaving room for technological improvement in the future.

We integrated the code of the face-api.js library with a freely and publicly available JavaScript-based platform (Minno.js; Zlotnick et al., 2015) for running web studies. That platform was developed in our lab and has been used for almost a decade in the Project Implicit (Ratliff & Smith, 2021) website (https://implicit.harvard.edu), contributing to numerous published studies (e.g., Béna et al., 2022; Moran & Bar-Anan, 2018; Vaimberg et al., 2021). We activated the facial expression detection module every 30 ms in Study 1, and every 100 ms in Study 2. However, as we report below, the program did not provide that detection rate consistently across trials and across participants, leaving room for future technological development. When a specific facial expression (e.g., a smile) was consecutively the most probable

dominant expression for more than 400 ms, we considered that a facial expression response, and used JavaScript code to trigger an emulation of a key response. That programming solution allowed us to convert JavaScript code programmed for creating a key-response task into a facial-response task with minimal modifications to the code. To examine whether our technological implementation of the FR-IAT can facilitate research over the Internet, we administered our studies over the popular commercial online crowdsourcing platform Prolific (Palan & Schitter, 2018).

Overview of the studies

In each of the two studies reported in the main text, half of the participants completed a keyboard version of the IAT (i.e., the "original" IAT), and half completed the FR-IAT. We tested the internal consistency of the FR-IAT and its correlation with self-reported evaluation. We tested whether, beyond response latency, the intensity of the emotion (estimated from the mean probability that it is the dominant emotion, in a specific time range within the trial) is also related to evaluation. We also tested whether performance in Blocks 2 and 5 (those that do not include attribute categories) is related to evaluation. The attitude objects were *Cats* and *Dogs* (Study 1) and *Britain* and the *United States* (Study 2). In Study 2, we added a self-report evaluation measure for the specific photos used in the IATs, and explored whether performance in the FR-IAT is sensitive to the evaluation of the specific stimuli used as items of the target categories, rather than only to the evaluation of the target categories. In both studies, the purpose of the keyboard IAT condition was to estimate the psychometric qualities achieved in the same context with the IAT.

In Study 2, we also examined, graphically, the temporal dynamics of the responses in the FR-IAT within trials. Across participants, we plotted facial expression response as a function of (1) time from the trial onset and (2) the categorization condition. That plot provided a glimpse into the potential of the FR-IAT to provide rich data that are not available with the keyboard IAT (Study 1 was less suitable for such plots because the plots required clear compatible and incompatible categorization conditions, for each participant).

In the web supplement, we report two additional studies. In Study S1, we tested whether the FR-IAT would replicate results that were previously found with other indirect measures but not with self-reported evaluation. Such evidence would suggest that the mental processes that influence performance in the FR-IAT are more similar to the mental processes that influence other indirect measures than those that influence self-reported evaluation. Study S2 was conducted to estimate a delay in response latency measurement, introduced by a specific coding choice we made in our program, detected by a reviewer.

All the preregistrations, materials, data, and analysis code for these studies can be found at https://osf.io/7tyf9/. All the data processing decisions and statistical analyses reported here were as preregistered, unless explicitly stated otherwise.

Study 1

Participants

On Prolific, 196 English-speaking participants (52% women, $M_{age} = 29.8$, $SD_{age} = 10.4$) completed the study for payment (£3.75). We planned to recruit 200 participants, but four participants managed to complete the study twice, and their second session was excluded from the analyses. The main statistical tests pertained to the correlation between self-report measures and the scores of the IATs. Based on pilot studies, we expected to retain about 85 participants in each condition (FR-IAT vs. IAT), to achieve 80% power for detecting correlations of 0.3, and 97% for correlations of 0.4. We could not achieve sufficient power to detect small or medium differences between the correlations obtained with FR-IAT and those obtained with the IAT, because achieving 80% power for the detection of even a relatively large difference between a $r = 0.3$ and $r = 0.5$ requires about 550 participants, which is beyond our financial means. Therefore, we settled for testing whether the FR-IAT shows reasonable validity, and added the IAT condition for a rough estimation of the validity achieved with the original key-response measure.

All 89 participants who were assigned to the keyboard IAT and completed that task were eligible for data analysis, based on the IAT's recommended exclusion rules (Greenwald et al., 2003). Of the 104 participants who were assigned to the FR-IAT and completed that task, we excluded from the main analyses nine (9%) participants based on their performance in the task. We used the same exclusion criteria in all the studies. We excluded participants who did not provide enough scorable trials within each of the conditions that we used for scoring the FR-IAT. Specifically, a participant was excluded if, in at least one of the four parts relevant for the scoring (Block 2, Block 5, Blocks 3+4, or Blocks 6+7), in more than 30% of the participant's trials, our data processing did not detect a correct facial response before the response deadline, or detected such a response faster than 400 ms (which is probably too fast for a facial response; see Figs. 1 and 2). In the tables that summarize the results, we also provide information about the results without participant exclusion.

Materials

In the IATs, the category labels were *Cats* (items: sketch photos of cats), *Dogs* (items: sketch photos of dogs), *Bad words* (items: *Negative, Awful, Horrible, Horrific, Terrible*), and *Good words* (items: *Positive*, *Excellent*, *Fabulous*, *Fantastic*, *Wonderful*). In the facial expression practice task, we used 12 photos from the racially diverse affective expression (RADIATE) face stimulus set (Conley et al., 2018)—one smiling photo and one angry photo from three men and three women.

Design, procedure, and measures

Participants had a 57% probability of being randomly assigned to complete the FR-IAT, and 43% to the IAT condition. We programmed that imbalance due to a larger expected number of exclusions in the FR-IAT, based on pilot studies. The order of the pairing conditions in the IATs (i.e., whether Cats shared a response with *Good words* or with *Bad words* in Blocks 3 and 4) was counterbalanced between participants.

Facial expression practice After the consent form and general instructions, the study in both conditions started with a training task, unrelated to the IAT, designed to teach participants to respond facially. Even participants who completed the keyboard IAT started with that practice task, to reduce the differences between the two conditions.

Participants were instructed to "make a positive facial expression" when seeing a positive facial expression, and "make a negative facial expression" when seeing a negative facial expression. In Blocks 1 and 3, the ten randomly ordered trials showed five smiling and five angry faces. In Block 2, we asked participants to practice the negative response, and showed five trials of angry faces. We informed participants (based on pilot data) that our software does not always detect negative facial expressions well, and recommended that, if the software does not detect their angry face well, they should try showing their teeth, and even make a *grrr* sound. We added, "For some people, it is easier to show a sad of a disgusted face, so you can try showing sadness or disgust as your negative face."

Each trial started with a 300 ms fixation stimulus, replaced with the target face. The face appeared until a response, or until the response deadline (4 seconds). If the software did not detect a response within 4 seconds, a "!!!NO EXPRESSION DETECTED!!!" message appeared for 1.5 seconds. If the software detected an incorrect response, a red "!!!WRONG EXPRESSION!!!" message appeared for 1.5 seconds. At the end of the trial, the message "SHOW A NEUTRAL EXPRESSION" appeared and remained on the screen until detecting a neutral facial response, or after four seconds without detecting a response. Finally, a blank screen appeared for 250 ms, until the next trial began.

Questionnaire Next, participants answered six pairs of evaluation and evaluation-related questions. Within each pair of questions, one question pertained to dogs and the other pertained to cats (randomly ordered). The first pair of questions was *How much do you like cats[dogs]?* The next three pairs were randomly ordered: *How cold or warm are your feelings toward cats[dogs]? How negative or positive do you think that cats[dogs] are? Do you associate cats[dogs] with negative or positive concepts?* These four pairs of questions had a seven-point response scale. The questionnaire ended with two (randomly ordered) pairs of questions: *How cute are cats[dogs]?* and *How majestic are cats[dogs]?* These two pairs of questions had a five-point response scale.

FR-IAT We described the block sequence and the task in each block of the FR-IAT in the introduction of this article. The number of trials in the seven blocks was, respectively, 20, 28, 20, 32, 28, 20, 32, for a total of 180 trials. These numbers were slightly different than those of the typical keyboard IAT: we added eight trials to each block that required categorizing the categories (Blocks 2 and 5) because we suspected that performance in these blocks would be related to people's evaluation of the categories. We added the trials to those blocks to increase statistical power for detecting and estimating the compatibility effect. We were particularly interested in that effect because there was no prior research that tested the relation between categorizing stimuli to categories with emotional facial responses and liking those categories. To avoid exhausting the participants, we balanced that increment by removing eight trials from Blocks 4 and 7.

Each trial started with the display of the target stimulus. As in the typical keyboard IAT, we required participants to correct error responses: an error response triggered a red "WRONG! SHOW THE OTHER EXPRESSION TO CONTINUE" message. If the target stimulus appeared for 9 seconds with no detection of the correct response, a red "!!!NO EXPRESSION DETECTED!!!" message appeared for 1.5 seconds. After a correct response or a timeout message, the message "SHOW A NEUTRAL EXPRESSION" appeared and remained on the screen for 5 seconds or until a neutral facial expression was detected. Afterwards, a blank screen appeared for 250 ms before the next trial started.

The software attempted to detect emotional facial expressions every 30 ms. Table 2 presents the information provided by the software with each detection attempt. The output of each detection attempt was the probability, from 0 to 1, for each of seven emotions: happy, sad, angry, fearful, disgusted, surprised, neutral. The total of these seven probabilities was always 1. These seven numbers, for each emotion detection attempt, were the data saved to our server, along with the

Table 2 Example of data from the facial response detection module in a single trial

Timestamp	Time since trial began	Neutral	Happy	Sad	Angry	Fearful	Disgusted	Surprised	Negative (total)	Dominant
1650873141662	84	1	0	0	0	0	0	0	0	Neutral
1650873141781	119	1	0	0	0	0	0	0	0	Neutral
1650873141992	211	1	0	0	0	0	0	0	0	Neutral
1650873142262	270	1	0	0	0	0	0	0	0	Neutral
1650873142619	357	1	0	0	0	0	0	0	0	Neutral
1650873143059	440	1	0	0	0	0	0	0	0	Neutral
1650873143552	493	1	0	0	0	0	0	0	0	Neutral
1650873144106	554	1	0	0	0	0	0	0	0	Neutral
1650873144713	607	0.55	0	0	0.02	0	0.43	0	0.45	Neutral
1650873145403	**690**	0	0	0	0.01	0	0.99	0	1	Negative
1650873146153	750	0	0	0	0	0	1	0	1	Negative
1650873147018	865	0.05	0	0.07	0	0	0.88	0	0.95	Negative
1650873147946	928	0	0	0.01	0	0	1	0	1	Negative
1650873148915	969	0	0	0.01	0	0	0.99	0	1	Negative
1650873149889	974	0	0	0	0.01	0	0.99	0	1	Negative
1650873150938	1049	0	0	0	0.01	0	0.99	0	1	Negative
1650873152071	1133	0	0	0	0.01	0	0.99	0	1	Negative

In bold font, the response latency recorded for this trial. At 865 ms, the probability of the negative facial expression (computed as the sum of the probability of sad, angry, and disgusted) was the largest at least two consecutive times (in fact, three) and for more than 150 ms (in fact, 175). Therefore, our data processing algorithm determined that this trial had a negative facial response with a response latency of 690 ms (when the sequence of consecutively dominant negative expression started). When the task was administered, the same algorithm was used to detect a response, but required the response to be consecutively the most probable response for at least 400 ms. Therefore, the trial continued until 1133 ms; The column "negative" was computed as the sum of sad, angry, and disgusted; "happy" was the positive response

trial number and the timestamp of the detection (following the UNIX Epoch time convention, the timestamp was the number of milliseconds since 1 January 1970). To compute the likelihood of a negative emotional facial response in each detection, we used the sum of sad, angry, and disgusted (pilot data revealed that fear was not used by participants to convey a negative response). While the participant completed the task, our code determined that participant expressed an emotional facial response when the same emotion (smiling or a negative facial response) had the highest probability of all the emotions, for at least 0.4 seconds. This algorithm for detecting an emotional response was used only for administrating the task, to encourage the participants to show an emotion clearly for a reasonable time. We used a different threshold for detecting emotional facial responses in our data analysis (see below).

IAT The IAT had the same number of trials and the same block sequence as the FR-IAT, with one exception: As in the typical IAT, participants practiced sorting the items of the target categories in Block 1 (28 trials) and practiced sorting the items of the attribute categories in Block 2 (20 trials). The trial sequence was identical to the typical IAT (Greenwald et al., 2022): the target stimulus appeared from the beginning of the trial until the correct response, with no response deadline. An error response triggered the display of a red X, and remained until participants responded correctly. As in the typical IAT, in the combined-categorization blocks of the IAT and the FR-IAT, the target stimulus alternated each trial from attributes to categories (Greenwald et al., 2022).

Data processing

Questionnaire Because the IATs provided a preference score, we created, from each pair of questions, a difference score that reflected preference for dogs over cats. We did not aggregate these six preference scores, to allow multiple criteria for convergent validity tests. The range of correlations between these six scores was $.47 \leq r \leq .87$.

IAT We computed two IAT scores. One was a D600 score, one of the recommended scoring algorithms for the IAT (Greenwald et al., 2003): trials were excluded if the responses were faster than 300 ms or slower than 10 seconds. Latency of error trials was replaced by the mean latency of the participant in the trial's block + 600 (penalty). For each pair of blocks (3 and 6, 4 and 7) the D score was M1 − M2 / SD, where M1 is the mean of one block (e.g., Block 3), M2 is the mean of the other block (e.g., Block 4), and SD is the

standard deviation of the two blocks together. The IAT D score was the mean of these two D scores.

The second IAT score was a G score (Sriram et al., 2006), a scale-invariant, nonparametric dominance measure. For each participant, we first assigned fractional ranks (percentiles) to all N latencies not excluded from the relevant blocks (Blocks 3, 4, 6, and 7). We then subtracted 1/2N from each fractional rank. We next standardized the ranks (i.e., computed the standard normal deviate, with mean = 0 and standard deviation = 1). The G score was the difference between the mean standardized ranks of the two conditions (i.e., when *Cats* shared a response with *Good words* vs. when *Dogs* shared a response with *Good words*). Previous research has found that the G score is not inferior to the D score (Richetin et al., 2015). We used the G score in the present research because it was easily adapted for the FR-IAT, in which the performance variable was not always latency (see below). As shown in Tables 3 and 4, the psychometric quality of the G score for the IAT was similar to the psychometric quality of the IAT D score.

FR-IAT The rich data collected with the FR-IAT would require extensive research for determining the best data processing and scoring algorithms. For this report, we relied on scoring algorithms that seemed reasonable to us, and showed a reasonable performance in pilot studies. Notably, however, different algorithms were superior in different pilot studies, suggesting that the best algorithm has yet to be found. To increase the confidence in the results, we focus in this article on the scoring algorithm and exclusion rules that we preregistered. As an alternative scoring method, we added a method based on the findings of the within-trial analyses (see more below).

We previously mentioned that when participants completed the task, our code determined that participant expressed an emotional facial response when the same emotion had the highest probability of all the emotions, for at least 0.4 seconds. We used the same logic for detecting responses in our data processing, only changing the duration required for consecutive dominance of an emotion from 400 ms to 150 ms. Across studies, using the 150 ms threshold provided satisfactory correct-response rate (see below), suggesting that it indeed detected participant's intended response. After detecting a response, the response latency was determined as the time that elapsed from the beginning of the trial and until the response started (see Table 2 for an example). Because the IATs required correction of error responses, the response latency of each IAT trial was the latency for the correct response.

We computed four scores, and used their mean as the focal score. Two of the four scores were based on response latency. Before computing the response latency scores, we removed trials in which the correct response was not detected at all (3.2% of the trials), and trials in which the correct response was detected in less than 400 ms (2.8% of the trials), which seems too quick for a facial expression response (see Figs. 1 and 2, for corroboration). The first response latency score was computed from the response

Table 3 Study 1: Descriptive statistics, internal consistency and correlations of the scores

	M (SD)	α	Liking	Positivity	Warmth	Association	Cute	Majestic
FR-IAT (planned scoring)								
Overall	−0.04 (0.37)	$.78_{a}$	**.47**	.39	**.45**	.43	**.45**	.25
$RT_{Blocks\ 3,4,6,7}$	−0.08 (0.48)	$.75_{abc}$	.43	.35	.39	.38	.44	.13
$RT_{Blocks\ 2,\ 5}$	0.00 (0.46)	$.48_{e}$	.35	.39	.40	.42	.36	**.34**
$Prob._{Blocks\ 3,4,6,7}$	−0.02 (0.38)	$.63_{bcde}$	.40	.29	.34	.32	.43	.13
$Prob._{Blocks\ 2,5}$	−0.04 (0.43)	$.61_{de}$	.38	.27	.37	.31	.29	.23
FR-IAT (alternative scoring)								
Overall	−0.04 (0.41)	$\mathbf{.79_{a}}$	.46	**.41**	**.45**	**.45**	**.45**	.26
$Prob._{Blocks\ 3,4,6,7}$	−0.07 (0.48)	$.77_{ab}$	.43	.34	.37	.37	.41	.16
$Prob._{Blocks\ 2,5}$	0.00 (0.45)	$.58_{bcde}$	.39	.38	.40	.41	.38	.31
IAT								
D	−0.07 (0.48)	$.71_{abcd}$	.43	.39	.42	.36	.44	.33
G	−0.10 (0.52)	$.68_{abcd}$	**.47**	.39	.44	.38	.44	.33
No exclusions								
FR-IAT (overall)	−0.04 (0.35)	$.75_{abc}$	.40	.43	.40	.39	.41	.23

Prob. probability-based; *RT* reaction time-based. The overall FR-IAT scores were the mean of the two latency-based scores and the two probability-based scores. The alternative FR-IAT scores were based on the mean probability of the correct response in the 500–1200 ms segment of the trials. Any correlation above .20 is statistically significant with $p < .05$. Any correlation above .25 is statistically significant with $p < .01$. Internal consistencies that do not share a subscript are significantly different. In **bold**: the largest value within each column. The last row shows the results of the FR-IAT overall score, without excluding any participant for poor performance

latency in the combined tasks blocks (Blocks 3, 4, 6, 7). Its computation was identical to how the IAT G score was computed. The second latency score was computed from the response latency in the blocks that required only categorizing the target categories (Blocks 2 and 5), and it followed the same logic (each block was one condition for computing the G score).

The other two scores that we computed from the FR-IAT data were based on the probability values provided within each trial, each time the software attempted to detect emotional facial expression (as shown in Table 2). We considered those probability values a proxy for the intensity of the facial expression. We used detection data from 100 ms to 2500 ms after the trial commenced. For each trial, we averaged, separately, the probability value for smiling and the probability value for a negative facial expression (the total of disgust + sadness + anger). For each trial, we then computed the difference between these two means, such that the difference always reflected by how much the mean of the correct facial expression in that trial was larger than the mean of the incorrect facial expression in that trial. We expected that probability difference to be larger, the easier the response was for the participant. That probability difference was the variable that we used to compute two G scores: one from the trials in the combined-categorization blocks, and one from the trials in the target categories blocks (Blocks 2 and 5).

After we conducted all the studies, and examined the effect of compatibility of the Block's categorization rules as a function of time since the beginning of the trial (see Figs. 1 and 2), we realized that, usually, the compatibility effect was apparent for a shorter segment of the trials. The compatibility between the liking of the categorized stimulus and the required response influenced performance most strongly in the period between 500 ms and 1200 ms after the onset of the target stimulus. Therefore, as exploratory alternative scores (not preregistered), we also computed two probability-based scores that used data collected in a segment between 500 ms and 1200 ms. We also computed an overall alternative FR-IAT score as the mean of the two response latency scores and that two probability-based scores in the range of 500–1200 ms. That alternative score usually performed better than the predetermined score, and probably can be improved further (e.g., by finding the most informative trial segment for each participant, separately).

Results and discussion

The FR-IAT data characteristics In the FR-IAT, we tried to detect the facial expression every 30 ms. However, the timestamp recorded with each detection revealed that the detection rate was often variable. For each participant, we computed the mean and standard deviation of the duration between detections. Ideally, all participants would show M = 30, and SD = 0. However, across participants, the distribution of the mean had $M = 116$ ms, $SD = 129$, and the distribution of the standard deviation had $M = 79$, $SD = 108$. This leaves much room for future technological improvement.

Across all trials in which participants responded correctly, participants were faster to smile ($M = 933$, $SD = 242$, $median = 859$) than to show a negative facial expression ($M = 1098$, $SD = 339$, $median = 1018$), $t(94) = 6.70$, $p < .001$, $d = 0.69$. Participants were also more likely to respond correctly when the correct response was smiling ($M = 0.97$, $SD = 0.05$, $median = 0.99$) than showing a negative response ($M = 0.94$, $SD = 0.07$, $median = 0.96$), $t(94) = 3.78$, $p < .001$, $d = 0.39$.

Evaluation The descriptive statistics and the relations between all the measures appear in Table 3. Both IATs showed a medium-strength relation with self-reported judgments of the pets. None of the correlations of any of the scores was significantly different from any correlation obtained with any other measure. The FR-IAT score that was based on response latency from the four blocks that combined attribute categorization with categorization of the target categories showed psychometric qualities very similar to the IAT score. All the additional scores that the FR-IAT can provide and the IAT cannot—those based on the probability of facial expression's dominance and those based on the blocks that required only the categorization of the target categories (Blocks 2 and 5)—were correlated with all the self-report measures, although their internal consistencies were lower than those of the scores based on response latency in the combined-categorization blocks. Integrating all four FR-IAT scores into an overall score did not result with a score that was much better, psychometrically, than the FR-IAT score that was based only on response latency in the combined-categorization blocks. In summary, performance in an FR-IAT designed to measure the preference between cats and dogs was related to people's reported preference between those animals. The FR-IAT was comparable to the keyboard IAT in its psychometric qualities.

Study 2

To further examine the validity of the FR-IAT as a performance-based indirect measure of evaluation, in a more social context, we designed an FR-IAT for the measurement of the evaluation of nations—Britain and the United States. We recruited participants who reported that they were located in one of those countries, to increase the likelihood that these attitude objects would be psychologically relevant for them.

Participants

On Prolific, 349 participants (47% women, M_{age} = 40.0, SD_{age} = 13.3), completed the study for payment (£4.05). We planned to recruit 350 participants, but the data for one participant were lost for technical reasons. As in the previous study, we planned to recruit enough participants for reasonable power to detect a small correlation among participants who completed the FR-IAT (specifically, 91% for $r = 0.25$), while acknowledging that we could not recruit a sample large enough to detect small differences between the FR-IAT and the IAT, in their correlation with other measures. Of the 192 participants who completed the FR-IAT, we excluded from the main analyses 25 (12%) participants based on their performance in the task, using the same exclusion rules as in Study 1. All 157 participants who completed the keyboard IAT were eligible for data analysis.

Materials

In the IATs, the attribute category labels were the same as in Study 1, and the target category labels were *Britain* (items: photos of Boris Johnson, Margaret Thatcher, the British flag, red double-decker busses, and the queen with the queen's guard) and the *United States* (items: photos of Donald Trump, Hillary Clinton, the US flag, the Statue of Liberty, and yellow taxis).

Design, procedure, and measures

The design and procedure was identical to Study 1, with the same facial expression sorting task for practice, and the following modifications for the evaluation measures.

Questionnaire The questionnaire started with five randomly ordered questions: two pairs of questions and one preference question: How much do you like *Britain [the United States]*? *How negative or positive are your feelings toward Britain [the United States]? Which do you prefer, Britain or the United States?* The next (randomly ordered) pair of questions was *How much do you feel that you are a part of Britain [the United States]*? Followed by the (randomly ordered) pair of questions *In international sports competitions, how sad or happy are you when someone from Britain [the United States] wins?* Next, participants answered *What is your main country of origin?* (response options: United States, Britain, other). Finally, we showed participants the ten (randomly ordered) photos from the IATs, and asked *How positive or negative are your feelings toward what this photo shows?* All the judgment questions had seven response options.

IATs Other than the stimuli, the only difference from Study 1's IATs was the number of trials. We used 20 trials in the attribute categorization block, 40 trials in each of the target category categorization blocks, and 30 trials in each of the combined-categorization blocks (Blocks 3, 4, 6, and 7). In the FR-IAT, instead of trying to detect emotional facial expressions every 30 ms, the between-detection duration was 100 ms. We hoped that a longer duration would be more suitable for the equipment of a larger number of participants, leading to a more stable and reliable detection.

Data processing

Questionnaires We created a difference score from the four pairs of questions, to correlate with the preference computed from the IATs. The range of correlations between these four scores was $.24 \leq r \leq .80$.

IAT and FR-IAT The focal scores of interest were for the preference between the two countries. For these scores, we followed the same scoring procedure as in Study 1, with higher scores reflecting stronger preference for Britain. We also computed an evaluation score for each specific photo. Because we did not plan in advance to test per-item scores, we did not preregister any plans regarding how to compute those scores and how to analyze them. In the IAT, we computed a photo's G score by comparing the latency in trials in which the category of the photo shared a key with *Good words* and trials in which the category of the photo shared a key with *Bad words*. Because these scores were based only on the combined blocks (30 photo trials in each condition), each of these scores relied on 2–4 trials for each photo in each pairing condition.

In the FR-IAT, we computed the photo's G score by comparing trials in which participants were required to smile in response to the photo and trials in which participants were required to show a negative response to the photo. Because those trials occurred in all the blocks (excluding the attribute categorization practice block), each of these scores relied on 5–8 trials for each photo in each pairing condition. For each photo, we computed a score based on response latency and a score based on estimated probability throughout the trial (in the range between 500 ms and 1200 ms), and averaged these two G scores into one photo evaluation score.

Results

The FR-IAT data characteristics Across participants, the distribution of the mean duration between detections had M = 110 ms, SD = 32, and the distribution of the standard deviation was M = 44, SD = 54. This is an improvement over the previous study, but still leaves room for improvement

in future technology. Across all trials in which participants responded correctly, participants were faster to smile (M = 892, SD = 163, *median* = 863) than to show a negative facial expression (M = 1034, SD = 285, *median* = 943), $t(166)$ = 7.91, $p < .001$, $d = 0.61$. Participants were also more likely to respond correctly when the correct response was smiling ($M = 0.97$, $SD = 0.04$, *median* = 0.98) than showing a negative response ($M = 0.93$, $SD = 0.08$, *median* = 0.96), $t(166)$ = 4.67, $p < .001$, $d = 0.36$. These results were very similar to those found in Study 1.

Evaluation In Table 4, we present the descriptive statistics, the relation between the IAT scores and self-reported judgment, and the effect of reported nationality on the score (the latter effect was computed only from participants whose reported nationality was one of the two target countries). Numerically, the FR-IAT focal score had a stronger correlation with all the self-report measures, in comparison to the keyboard IAT score, whereas the IAT score showed a stronger sensitivity to self-reported nationality. However, none of the differences was statistically significant. As in Study 2, the scores computed from performance in the blocks that required only categorizing the target categories were related to self-reported evaluation, with numerically weaker relations than the scores based on the performance in the combined-categorization blocks.

We explored the sensitivity of performance in the FR-IAT to evaluation of each individual stimulus presented in the task by computing the correlation of the self-reported feelings toward each photo and all the per-item evaluation scores computed from the IATs. The left part of Table 5 shows the correlation between the self-reported evaluation of each photo and the same photo's evaluation score in the IATs. For all ten photos, the self-reported evaluation was more strongly related to the FR-IAT score than to the IAT score. Five of the ten FR-IAT item scores showed a correlation above 0.3, in comparison to only one of the ten IAT item scores. In other words, the FR-IAT's per-item evaluation scores had a stronger convergent validity than the IAT's per-item evaluation scores.

To evaluate the discriminant validity of the per-photo scores, we compared the correlation between self-reported judgment of each photo and the performance-based score computed for that specific photo with the correlations of the same self-reported judgment with all the other photo-specific scores. For that, we computed, for each photo, the mean absolute correlation of the self-reported evaluation of that photo with the evaluation scores of each of the other nine photos. We used the absolute correlation because we did not expect participants who liked one photo to also like another photo. If a particular photo evaluation score has good discriminant validity, then the correlation of that score with the self-reported evaluation of that photo would be stronger than the mean absolute correlation of the self-reported evaluation of the photo and the evaluation scores of all other photos. The results appear in Table 5. In eight out of the ten stimuli,

Table 4 Study 2: Descriptive statistics, internal consistency, and correlations

	M (SD)	α	National-ity effect	Liking	Positivity	Preference	Belonging	Sport Support	By-item prefer-ence
FR-IAT(planned score)									
Overall	0.01 (0.41)	$.84_{a}$	1.88	.40	.37	.58	.72	**.59**	.32
$RT_{Blocks\ 3,4,6,7}$	0.01 (0.52)	$.84_{a}$	1.94	.36	.33	.59	.69	.55	.25
$RT_{Blocks\ 2,\ 5}$	0.00 (0.52)	$.70_{cd}$	1.45	.38	.34	.49	.62	.52	.33
$Prob._{Blocks\ 3,4,6,7}$	0.00 (0.41)	$.75_{bcd}$	1.66	.32	.30	.52	.65	.55	.23
$Prob._{Blocks\ 2,5}$	0.03 (0.42)	$.67_{d}$	1.09	.35	.34	.44	.55	.45	.33
FR-IAT (alternative score)									
Overall	0.03 (0.42)	$\mathbf{.87}_{a}$	2.00	**.42**	**.38**	**.60**	**.73**	**.59**	.33
$Prob._{Blocks\ 3,4,6,7}$	−0.02 (0.51)	$.86_{ab}$	1.99	.38	.35	.58	.70	.57	.26
$Prob._{Blocks\ 2,5}$	0.00 (0.50)	$.73_{cd}$	1.45	.41	.37	.52	.63	.51	**.34**
IAT									
D	0.07 (0.51)	$.80_{abc}$	1.88	.28	.21	.54	.66	.54	.17
G	0.11 (0.60)	$.80_{abc}$	**2.03**	.32	.22	.58	.68	.58	.16
No exclusions									
Overall	0.01 (0.40)	.83	1.78	.39	.35	.56	.69	.57	.30

By-item preference: the preference computed from the ratings of the photos. All correlations are statistically significant with $p < .05$. Any correlation above .22 is statistically significant with $p < .01$. Internal consistencies that do not share a subscript are significantly different. The last row shows the results of the FR-IAT overall score, without excluding any participant for poor performance. In **bold**: the largest value within each column

Table 5 Study 2: Convergent and discriminant validity for single-item evaluation scores

Item	Correlation with self-reported evaluation of the item		Mean absolute correlation with all the other items		Difference (discriminant validity)	
	IAT	FR-IAT	IAT	FR-IAT	IAT	FR-IAT
Boris Johnson	−.04	.14	.14	.15	−.17	.01
Margaret Thatcher	.02	.13	.13	.08	−.11	.05
Red bus	.00	.29	.07	.15	−.06	.15
The queen's guard	.10	.37	.06	.18	.04	.19
British flag	.11	.33	.08	.16	.03	.17
Donald Trump	.33	.37	.06	.10	.27	.27
Hillary Clinton	.20	.25	.07	.10	.13	.16
Yellow taxi	.11	.12	.06	.07	.05	.04
Statue of Liberty	.17	.33	.15	.12	.02	.22
US flag	.29	.37	.16	.14	.14	.23

Convergent validity was tested as the correlation with the self-reported evaluation of the items; To test for discriminant validity, we compared the convergent validity correlations with the mean absolute correlation of the self-reported evaluation of each item with the evaluation scores of each of the other items. The difference between these correlations appears on the right-side columns—more positive numbers reflecting better discriminant validity

the discriminant validity of the FR-IAT was higher than that of the keyboard IAT. However, the probability of such an 8:2 ratio when the expected ratio is 5:5 is 5.7%, which does not pass the traditional 5% threshold for significant results.

Within-trial information The FR-IAT provides data a few times within each trial. These data can help investigate the response dynamics within trials. To demonstrate trial dynamics information, in Fig. 1, we plotted the mean estimated probability that the participant is showing the correct facial response, as a function of the correct response (smiling or showing a negative facial response), the duration since the stimulus was shown, and the compatibility of the current pairing condition with the participant's assumed preference between the nations. The duration's time unit was rounded to units of 0.1 second. The compatibility was computed for each participant based on their reported country of origin. Therefore, we excluded from this analysis participants who reported that their main country of origin was not the United States or Britain. We also excluded trials with incorrect responses. The mean probability was the mean of the estimated probability of the relevant emotional facial response at that time point, across all eligible trials from all the participants. We used data from the combined-categorization blocks (Blocks 3, 4, 6, and 7). As explained later, the time estimates are not exact because they include the time until the browser received the video frame from the webcam and the time of processing the video frame into emotions.

From the plot, it appears that the software detects smiling more easily than a negative response (larger probability most of the time). The compatibility effect on estimated probability of the correct response started after 500 ms, and became larger until peaking at 900 ms. Then, the compatibility effect decreased and disappeared at about 1200 ms. Figure 2 shows the same plot, based on trials from the Blocks 2 and 5 (when participants categorized only the target categories). The pattern of compatibility effects was almost identical to the pattern observed in the combined-categorization blocks.

Discussion

The results of Study 2 validated further the FR-IAT as a performance-based measure of evaluations. The preference scores computed from the FR-IAT were related to self-reported evaluation at least as strongly as the keyboard IAT. We also found initial evidence that the FR-IAT might be useful for computing evaluation scores of single stimuli.

Additional studies

In Studies 1 and 2, we measured the evaluation of objects that are expected to elicit similar evaluation spontaneously and intentionally. People seldom need to hide or alter their evaluation of pets or nations. There is no social pressure to reject one's spontaneous preference between pets or nations. These categories helped us validate the FR-IAT as a performance-based measure of evaluation, as typically done in validation research of indirect evaluation measures (e.g., Bluemke & Friese, 2008; Payne et al., 2005). However, indirect measures of evaluation are often used to capture evaluation that is unintentional and not easily controlled by motivation to conceal or correct one's spontaneous evaluation.

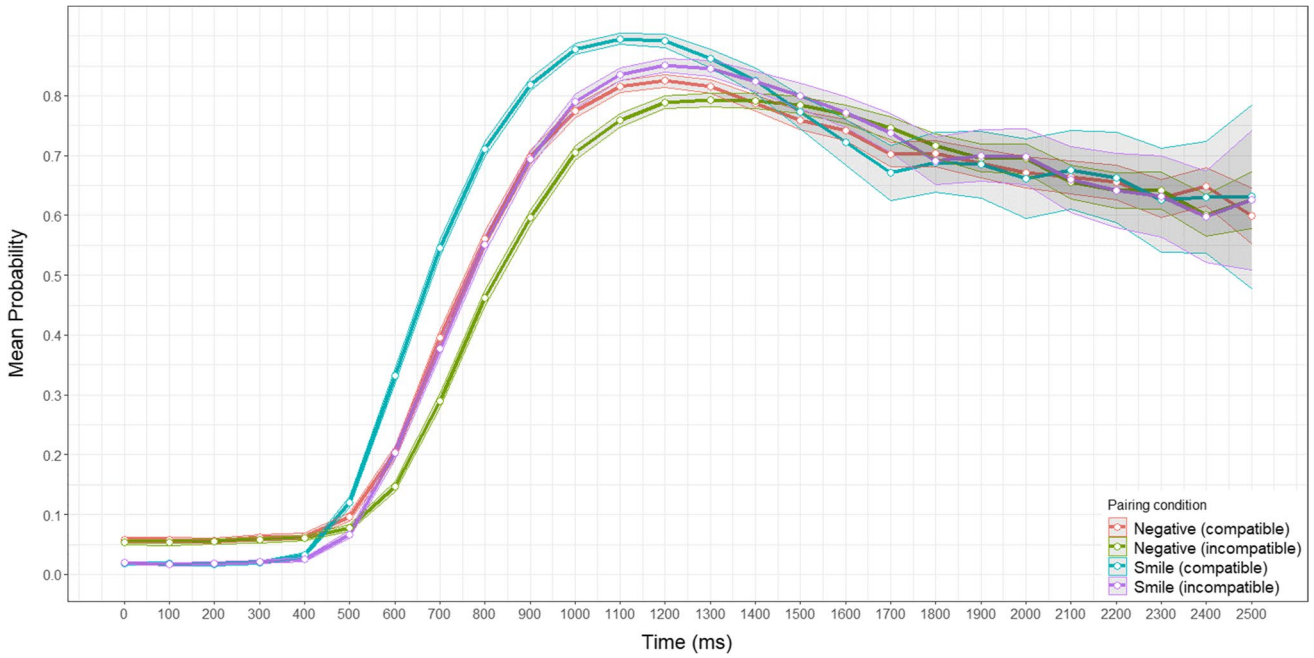

Fig. 1 Combined-tasks blocks in Study 2: the mean probability of the correct facial response, averaged across all correct-response trials of all participants, as a function of time since target stimulus onset, the correct facial response, and the compatibility of the pairing condition with the participant's assumed preference (inferred from the participant's reported main country of origin). In gray ribbons: confidence intervals. As explained in the main text, the recorded time in the *x*-axis included the time of processing the captured video frame into emotional expressions. According to the results of Study S2 in the web supplement, that might add about 100 ms to the time in which the facial response was actually captured

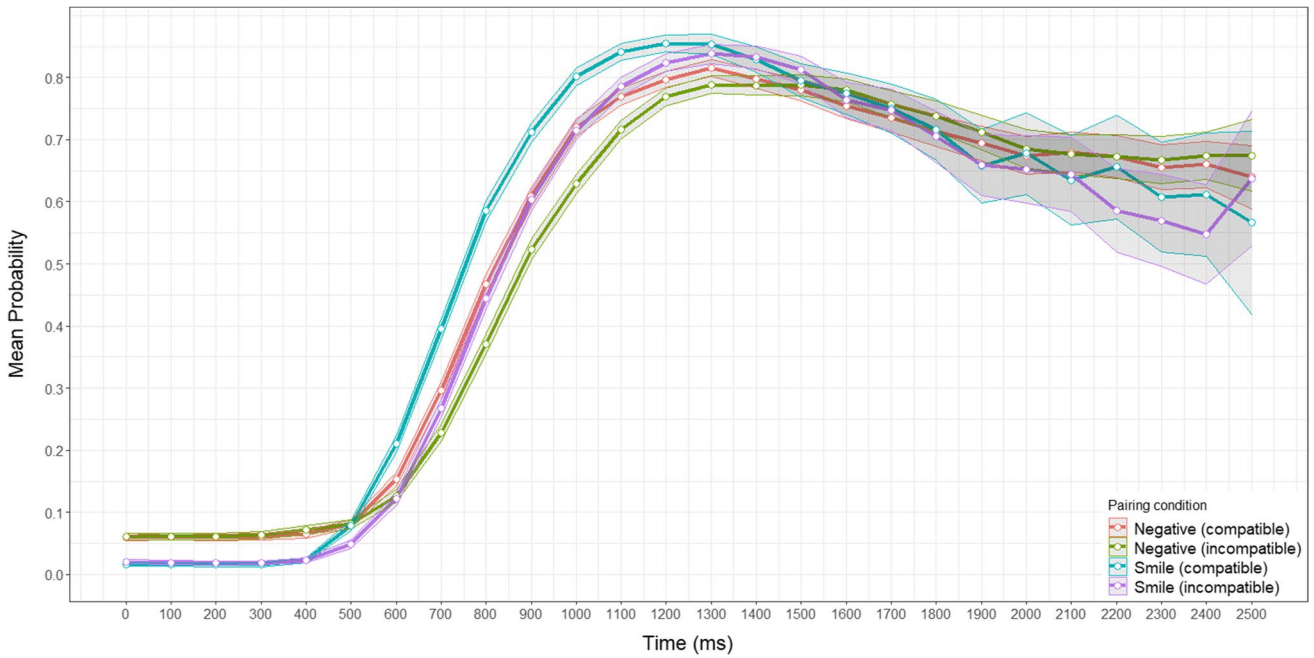

Fig. 2 Blocks 2 and 5 in Study 2: the probability of the correct facial response, averaged across all correct-response trials of all participants, as a function of time since target stimulus onset, the correct facial response, and the compatibility of the pairing condition with the participant's assumed preference (inferred from the participant's reported main country of origin). In gray ribbons: confidence intervals. As explained in the main text, the recorded time in the *x*-axis included the time of processing the captured video frame into emotional expressions. According to the results of Study S2 in the web supplement, that might add about 100 ms to the time in which the facial response was actually captured

To test whether the FR-IAT captures unintentional evaluation, we conducted Study S1, which is detailed in the online supplement.

In Study S1, we used a paradigm that previously found consistent discrepancy between self-reported evaluation and indirectly measured evaluation. Specifically, recent research in our lab (Navon & Bar-Anan, 2023) found that after reading about a man who is characterized mostly by positive behaviors and a woman who is characterized mostly by negative behaviors, female participants' performance in two indirect measures of evaluation—the (keyboard) IAT and the Evaluative Priming Task (EPT; Fazio et al., 1986)—reflects a preference for the woman over the man, despite self-reported preference for the man. That discrepancy can be interpreted as a bias in favor of the women's own gender group that occurs only when participants do not control their evaluative response.

Study S1 replicated the results previously found with the IAT and EPT: female participants reported a preference for the positive male target but none of the FR-IAT scores reflected such a preference, and almost all of them suggested a preference for the negative female target over the positive male target. That result is incompatible with the possibility that the FR-IAT is more sensitive to controlled evaluation than the IAT, and increases the likelihood that, like the IAT (and EPT), the FR-IAT is sensitive to unintentional evaluation. On the other hand, it is not yet well established that the discrepancy between self-reported evaluation and indirectly measured evaluation in Study S1 and in the studies it replicated indeed reflects the discrepancy between intentional and unintentional evaluation. Because the validation of the IAT as a measure of unintentional evaluation is still a matter of active research (Vianello & Bar-Anan, 2021), we cannot rule out the possibility that the discrepancy reflects a shared sensitivity of the IAT, EPT, and the FR-IAT to nonevaluative factors that do not influence self-reported evaluation.

We conducted another auxiliary study (Study S2) to address the consequence of a programming choice in our code, detected by a reviewer. Our code recorded the time of the facial expressions captured by the participant's camera only after the JavaScript library processed and detected the emotional expressions. That introduced a measurement delay because detection was not immediate. In Study S2, we computed the mean delay (i.e., the duration of processing the visual input into emotional facial responses) for each participant ($M = 108$; $SD = 30$). That means that the times recorded in all the other studies were actually shorter. For example, Figs. 1 and 2 show that the onset of the emotional facial response was typically at about 400 ms, but that was probably an overestimation (by about 100 ms). We also computed the mean standard deviation of the duration of processing the visual input into emotional expressions, within each participant ($M = 25$; $SD = 28$). That variability was measurement noise that added inaccuracy to the computation of differences between different task conditions, for each participant. Improving our code to decrease that noise might improve the overall measurement quality, although such improvement would probably be negligible (Brand & Bradley, 2012; Damian, 2010; Ulrich & Giray, 1989).

General discussion

In this article, we presented the Facial Response Implicit Association Test (FR-IAT)—a variant of the IAT that replaces key responses with facial responses. To administer the FR-IAT over the Internet, we integrated a publicly available JavaScript library for detection of emotional facial responses through webcams, with a publicly available JavaScript software for creating and executing behavioral research over the Internet. We found evidence in favor of the FR-IAT as a useful indirect measure of evaluation. Performance in the FR-IAT was related to self-reported evaluation with correlations that were very similar to the correlations that we found with the keyboard IAT. Numerically, the internal consistency and correlations of the FR-IAT were usually superior to those found with the IAT, although the differences were small and never reached statistical significance.

In Study 2, we found that evaluation FR-IAT scores computed for each specific stimulus often correlated with self-reported evaluation of that stimulus more than with self-reported evaluation of any other stimulus. That evidence for convergent and discriminant validity of the item-specific evaluation scores was often stronger than the evidence found for the item-specific evaluation scores computed from the keyboard IAT. Further improvement of this potential strength can provide measurement features that the current keyboard IAT lacks. Because measuring evaluation toward the specific items is not always a priority for researchers, it is noteworthy that we found no evidence that this sensitivity harmed the measurement of the evaluation of the categories in the FR-IAT, in comparison to the IAT.

One obvious reason for the FR-IAT's potential superiority in measuring single-item evaluation is that unlike in the keyboard IAT, the trials of almost all the FR-IAT's blocks (excluding the block that provides practice in categorizing attribute stimuli) are sensitive to evaluation. In all the studies, in the blocks that required only categorization of the items that belong to the target categories (Blocks 2 and 5 of the FR-IAT), it was easier for participants to smile toward items of their favorite category and show a negative facial expression toward items of the disfavored category. In other words, because the responses in the FR-IAT are strongly related to evaluation, the level of performance in categorizing stimuli to target categories may be influenced by the person's evaluation of the category and the categorized

stimulus. The present research found evidence in support of that possibility. Notably, however, the scores computed from the FR-IAT's combined-tasks blocks (when attribute and target categories share a response) were always numerically superior to the scores computed from the category-only blocks.

Opportunities for improvement

The present results are especially encouraging when considering the possible opportunities for improvement of the FR-IAT. First, at this early stage, we have not yet taken advantage of the data provided about the facial responses throughout each trial. It is still unclear how evaluative processes influence the response dynamic within each trial. For example, we have not yet examined the relation between evaluation and the pattern of changes of the response as a function of time (e.g., the slope). We hope that the present data will facilitate future research on the best methods for detecting evaluation based on performance in the FR-IAT. There might also be room for improving the task procedures. We already mentioned that in our pilot studies we found that our software detected a negative response better when we replaced the instructions to frown with detailed instructions to show any negative facial reaction. It is likely that further modifications of the instructions would improve the measure. For example, perhaps requiring participants to respond with a disgusted facial expression would improve the task and its measurement quality. That could be the case because disgust might be more strongly related to negative evaluation than anger or sadness, or due to technical reasons related to the ease of producing disgusted expressions or the ease of detecting those expressions automatically. Other adaptations of the FR-IAT may serve research with special populations that struggle with keyboard responding and may find it easier to respond with their facial expressions (e.g., young children or the elderly).

Another avenue for improvement of the FR-IAT is to use facial expression detection software that would provide richer information than the software that we used in the present research. Existing computer programs for video analysis (e.g., openFace, Baltrusaitis et al., 2018) already provide rich data about the intensity of each facial muscle contraction, sometimes with idiosyncratic adjustments for the personal characteristic of each videotaped individual. It could be possible to enrich the FR-IAT's data by videotaping participants while they complete the task. However, a more useful solution would be to develop JavaScript code that supports some of those more advanced capabilities. For example, even only differentiating between genuine and non-genuine smiles, or between different intensities of the emotional facial expressions, may improve the measurement quality. At the same time, there is also room for improving the detection of basic emotions to reduce the likelihood of losing data because the facial expressions of some people (e.g., those with facial hair or glasses) are not easily detected. It may also be possible to improve the performance of the JavaScript code, such that the duration between detections would become less variable between and within participants.

In Blocks 2 and 5 of the FR-IAT, evaluation was measured by comparing the participant's performance when a positive facial expression (smiling) was the correct response to the target object with the participant's performance when a negative facial expression was the correct response to the same object. Would it be useful to simplify the FR-IAT by using only those conditions for measurement? In the present research, scores based only on performance in Blocks 2 and 5 showed reasonable validity, although slightly inferior (numerically) to the validity evidence that we found for the FR-IAT's combined-categorization blocks. In fact, a previous study found evidence suggesting that the evaluation of the target objects influences facial expression categorization even when the categorization task (based on whether the photo is tilted or upright) is unrelated to the target categories (exercising vs. sedentary office work), and the target categories are never mentioned in the task (Brand & Ulrich, 2019). Future research could test the psychometric qualities of such simple tasks, to examine and quantify the benefits of using the more complex FR-IAT. It would also be informative to investigate whether adding facial responses to other existing indirect measures, such as the EPT, the affect misattribution procedure (Payne et al., 2005), and the Brief-IAT (Sriram & Greenwald, 2009), would improve the psychometric quality of these measures.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02060-1.

Funding This work was supported by a grant from the Israel Science Foundation (ISF; grant No. 1684/21) to Y.B.A.

References

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

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 59-66). IEEE.

Banaji, M. R., & Heiphetz, L. (2010). Attitudes. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, 5th ed., pp. 353–393). Wiley.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Bavelas, J. B., & Chovil, N. (1997). Faces in dialogue. In J. A. Russell & J. M. Fernandez-Dols (Eds.), *The Psychology of Facial Expression* (pp. 334–346). Cambridge University Press.

Béna, J., Melnikoff, D. E., Mierop, A., & Corneille, O. (2022). Revisiting dissociation hypotheses with a structural fit approach: The case of the prepared reflex framework. *Journal of Experimental Social Psychology, 100*, 104297.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977–997.

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of Web experiments measuring response times. *Social Science Computer Review, 30*(3), 350–357.

Brand, R., & Ulrich, L. (2019). I can see it in your face. Affective valuation of exercise in more or less physically active individuals. *Frontiers in Psychology, 10*, 2901.

Brown-Iannuzzi, J. L., Cooley, E., McKee, S. E., & Hyden, C. (2019). Wealthy Whites and poor Blacks: Implicit associations between racial groups and wealth predict explicit opposition toward helping the poor. *Journal of Experimental Social Psychology, 82*, 26–34. https://doi.org/10.1016/j.jesp.2018.11.006

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure? *Wiley Interdisciplinary Reviews: Cognitive Science, 10*(5), e1501.

Carpenter, T. P., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S., LaBouff, J., Isenberg, N., & Chakroff, A. (2019). Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods, 51*(5), 2194–2208.

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. J. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research, 270*, 1059–1067.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods, 42*(1), 205–211.

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

Dimberg, U., Thunberg, M., & Grunedal, S. (2002). Facial reactions to emotional stimuli: Automatically controlled emotional responses. *Cognition & Emotion, 16*(4), 449–471.

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

Fernandez-Dols, J. M., Sanchez, F., Carrera, P., & Ruiz-Belda, M. A. (1997). Are spontaneous expressions and emotions linked? An experimental test of coherence. *Journal of Nonverbal Behavior, 21*(3), 163–177.

Field, M., Eastwood, B., Bradley, B. P., & Mogg, K. (2006). Selective processing of cannabis cues in regular cannabis users. *Drug and Alcohol Dependence, 85*(1), 75–82.

Friese, M., Hofmann, W., & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology, 47*(3), 397–419. https://doi.org/10.1348/014466607X241540

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). Cambridge University Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 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*(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

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

Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology, 43*(3), 497–504. https://doi.org/10.1016/j.jesp.2006.05.004

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Toward an individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology, 95*(4), 962–977. https://doi.org/10.1037/a0012705

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. *American Psychologist, 74*(5), 569–586. https://doi.org/10.1037/amp0000364

Messinger, D. S., Fogel, A., & Dickson, K. L. (1999). What's in a smile? *Developmental Psychology, 35*(3), 701.

Moran, T., & Bar-Anan, Y. (2018). How actions change liking: The effect of an action's outcome on the evaluation of the action's object. *Journal of Experimental Psychology: General, 147*(11), 1597.

Navon. M., & Bar-Anan, Y. (2023). The effect of individuating information and group membership on the deliberate and automatic evaluation of novel individual members of known social groups. *Manuscript in Preparation.*

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., Cunningham, W. A., Banaji, M. R., & Greenwald, A. G. (2000). *Measuring implicit attitudes on the Internet*. Society for Personality and Social Psychology.

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*(2), 166–180.

Osgood, C. E. (1962). Studies on the generality of affective meaning systems. *American Psychologist, 17*, 10–21.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parkinson, B. (2005). Do facial movements express emotions or communicate motives? *Personality and Social Psychology Review, 9*(4), 278–311.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293.

Ratliff, K. A., & Smith, C. T. (2021). Lessons from two decades of Project Implicit. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge handbook of implicit bias and racism*. Cambridge University Press.

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling

Implicit Association Test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. *PLoS One, 10*(6), e0129601.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*(2), 139.

Russell, J. A., Bachorowski, J. A., & Fernández-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*(1), 329–349.

Schneider, I. K., & Mattes, A. (2021). Mix is different from nix: Mouse tracking differentiates ambivalence from neutrality. *Journal of Experimental Social Psychology, 95*, 104106.

Smeding, A., Quinton, J. C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. *Journal of Personality and Social Psychology, 111*(6), 817–834.

Smith, C. T., & Ratliff, K. A. (2015). Implicit measures of attitudes. In T. Ortner & F. van den Vijver (Eds.), *Behavior Based Assessment in Psychology: Going Beyond Self-Report in the Personality, Affective, Motivation, and Social Domains* (pp. 113–132). Hogrefe.

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

Sriram, N., Nosek, B. A., & Greenwald, A. (2006). Scale invariant contrasts of response latency distributions. SSRN. https://doi.org/10.2139/ssrn.2213910

Uhlmann, E. L., Brescoll, V. L., & Paluck, E. L. (2006). Are members of low status groups perceived as bad, or badly off? Egalitarian negative associations and automatic prejudice. *Journal of Experimental Social Psychology, 42*(4), 491–499. https://doi.org/10.1016/j.jesp.2004.10.003

Ulrich, R., & Giray, M. (1989). Time resolution of clocks: Effects on reaction time measurement—Good news for bad clocks. *British Journal of Mathematical and Statistical Psychology, 42*(1), 1–12.

Vaimberg, E., Demers, L., Ford, E., Sabatello, M., Stevens, B., & Dasgupta, S. (2021). Project Inclusive Genetics: Exploring the impact of patient-centered counseling training on physical disability bias in the prenatal setting. *PLoS One, 16*(8), e0255722.

Vanman, E. J., Saltz, J. L., Nathan, L. R., & Warren, J. A. (2004). Racial discrimination by low-prejudiced Whites: Facial movements as implicit measures of attitudes related to behavior. *Psychological Science, 15*(11), 711–714.

Vanman, E. J., Ryan, J. P., Pedersen, W. C., & Ito, T. A. (2013). Probing prejudice with startle eyeblink modification: A marker of attention, emotion, or both. *International Journal of Psychological Research, 6*, 30–41.

Vianello, M., & Bar-Anan, Y. (2021). Can the Implicit Association Test measure automatic judgment? The validation continues. *Perspectives on Psychological Science, 16*, 415–421.

Voß, A., Rothermund, K., & Wentura, D. (2003). Estimating the valence of single stimuli: A new variant of the affective Simon task. *Experimental Psychology, 50*(2), 86–96.

Wiers, R. W., Beckers, L., Houben, K., & Hofmann, W. (2009). A short fuse after alcohol: Implicit power associations predict aggressiveness after alcohol consumption in young heavy drinkers with limited executive control. *Pharmacology, Biochemistry, and Behavior, 93*(3), 300–305. https://doi.org/10.1016/j.pbb.2009.02.003

Zlotnick, E., Dzikiewicz, A., & Bar-Anan, Y. (2015). Minno. js (Version 1.0)[Computer software]. https://minnojs.github.io/

Open practices statement All data, preregistration documents, and materials for all three studies are available at https://osf.io/7tyf9/.