Received: 22 April 2021 | Accepted: 28 September 2023

DOI: 10.1111/bjso.12688

REGISTERED REPORT STAGE 2

Implicit association tests: Stimuli validation from participant responses

Sally A. M. Hogenboom[1,2] | Katrin Schulz[1] | Leendert van Maanen[3]

[1]Faculty of Humanities, Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Theory, Methods, and Statistics, Faculty of Psychology, Open Universiteit, Heerlen, The Netherlands

[3]Department of Experimental Psychology, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, The Netherlands

Correspondence
Sally A. M. Hogenboom, Department of Theory, Methods, and Statistics, Faculty of Psychology, Open Universiteit, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands.
Email: sally.hogenboom@gmail.com

Funding information
Dutch Research Council (NL), Grant/Award Number: 406.DI.19.059

Abstract

The Implicit Association Test (IAT, Greenwald et al., *J. Pers. Soc. Psychol.*, *74*, 1998, 1464) is a popular instrument for measuring attitudes and (stereotypical) biases. Greenwald et al. (*Behav. Res. Methods*, *54*, 2021, 1161) proposed a concrete method for validating IAT stimuli: appropriate stimuli should be familiar and easy to classify – translating to rapid (response times <800 ms) and accurate (error < 10%) participant responses. We conducted three analyses to explore the theoretical and practical utility of these proposed validation criteria. We first applied the proposed validation criteria to the data of 15 IATs that were available via Project Implicit. A bootstrap approach with 10,000 'experiments' of 100 participants showed that 5.85% of stimuli were reliably valid (i.e., we are more than 95% confident that a stimulus will also be valid in a new sample of 18- to 25–year-old US participants). Most stimuli (78.44%) could not be reliably validated, indicating a less than 5% certainty in the outcome of stimulus (in)validity for a new sample of participants. We then explored how stimulus validity differs across IATs. Results show that only some stimuli are consistently (in)valid. Most stimuli show between-IAT variances, which indicate that stimulus validity differs across IAT contexts. In the final analysis, we explored the effect of stimulus type (images, nouns, names, adjectives) on stimulus validity. Stimulus type was a significant predictor of stimulus validity. Although images attain the highest stimulus validity, raw data show large differences within stimulus types. Together, the results indicate a need for revised validation criteria. We finish with practical recommendations for stimulus selection and (post-hoc) stimulus validation.

KEYWORDS
implicit association test, internal validity, stimulus validation

INTRODUCTION

The *Implicit Association Test* (IAT, Greenwald et al., 1998) is a popular measurement of implicit attitudes and (stereotypical) biases. New IATs are continuously created, and existing IATs are adapted and/or used in new experimental designs. Indeed, within the last 4 years, the number of IAT studies increased from 3608 (March 2019, Greenwald et al., 2020) to 4473 (August 2023): an average of 16 publications per month.[1]

Some of these publications raise concerns about the construct validity of the IAT (cf., De Houwer, 2001; Gawronski, 2009). For example, studies show low predictive validity (Greenwald et al., 2009), low test–retest reliability (Hehman et al., 2019), and a lack of discriminant validity (Schimmack, 2021). A possible explanation for the IAT's measurement issues is the lack of convergence in the utilized stimuli. Axt et al. (2021) argued that stimulus variability has the potential to degrade measurement quality, limit generalizability, cause misinterpretation of (null-) results, and affect associations with other measures. Detrimental effects, such as those argued by Axt et al. (2021), are not only theoretical. Multiple studies show that stimulus choices directly affect the size and direction of the measured IAT bias (Bluemke & Friese, 2006; Steffens & Plewe, 2001). Few articles, however, address how to prevent these measurement issues by selecting appropriate and valid stimuli. A notable exception is an article by Greenwald et al. (2021), which offers practical guidelines for designing and administering IATs. Concerning stimulus selection, the authors propose that the included stimuli should be familiar and easy to classify – translating to rapid (response times <800 ms) and accurate (error < 10%) participant responses (Greenwald et al., 2021, p. 7). As these criteria have only recently been published, empirical studies that evaluate and implement these guidelines have not yet been conducted. In the present study, we thus explored the theoretical and practical utility of these criteria as validation measures for IAT stimuli.

The implicit association test

The IAT measures implicit attitudes and (stereotypical) biases in terms of association strengths between categories. The *Gender-Career IAT* (GC-IAT), for example, is aimed at understanding implicit attitudes towards traditional gender roles by measuring association strengths between the categories Career/Family and Male/Female. In the present section, we specifically discuss the IAT's stimuli. A more detailed description of the entire IAT paradigm is available in Appendix A.

The IAT consists of categories and exemplars – together called the stimuli. The categories are labels that refer to, for example, social groups (e.g., "Christian" vs. "Muslim," Heiphetz et al., 2013) or attitudes (e.g., "Pleasant" vs. "Unpleasant," Greenwald et al., 1998). An IAT contains two sets of opposing categories $N_{\text{total}} = 4$, which together form the IAT's areas of interest. Each of these four categories in an IAT is represented by multiple exemplars: nouns, names, adjectives, images, and more. For participants, the objective is to sort the exemplars into the correct categories by pressing the corresponding keyboard keys. Central to our research is the fact that both the categories and exemplars exert an effect on the IAT's outcome measure D_{IAT} (Gast & Rothermund, 2010) and (in)appropriate stimuli selection, as we will discuss in the section below, directly affects the (direction of) the measured bias score.

Undesirable stimulus effects

Although it is evident that the IAT is only as good as the included stimuli – Greenwald et al. (preprint: 2020; article: 2021) are the only researchers to explicitly describe how stimuli should be selected

[1] We replicated the search strategy from Greenwald et al. (2020) on 28 August 2023. We conducted an advanced search on the American Psychological Association's PsycNET database (https://psycnet.apa.org/home) for publications including "Implicit Association Test" in the Title, Abstract, Keywords, *OR* Test & Measures.

to prevent undesirable stimulus effects. We focus specifically on undesirable stimulus effects that occur due to an interaction between stimuli characteristics and participant characteristics: stimulus unfamiliarity and cross-category associations.

The issue of stimulus unfamiliarity

Greenwald et al. (2021) recommend that stimuli should be familiar to the participants (table 1: A1–A2). When participants are unfamiliar with categories, the IAT cannot be expected to measure bias, may cause spurious correlations, and yield negative biases or null-effects (Greenwald et al., 2020, 2021). Brendl et al. (2001) even showed that unfamiliar non-word stimuli elicited more negative biases than familiar negative words. The need for familiarity however does not apply as strictly to the exemplars, nor does it apply when novel categories are first subjected to training (Greenwald et al., 2021).

(Un)familiarity with categories is largely dependent on the participant population. To illustrate, imagine a Hutu/Tutsi/Positive/Negative-IAT. Some participant populations (e.g., primary school students) may not be familiar with the labels “Hutu” and “Tutsi” describing two of the ethnic groups involved in the Rwandan genocide. Changing the population of interest (e.g., Rwandan vs. US students) may therefore induce stimulus unfamiliarity that did not exist before.

The issue of cross-category associations

A second important recommendation by Greenwald et al. is to select stimuli that avoid cross-category associations (Greenwald et al., 2021, table 1: A3–A7). Cross-category associations occur when the exemplar(s) of Category A can also be associated with Category B due to unintended stimuli and/or participant characteristics. For example, cross-category associations that exist because of stimuli characteristics are as follows: negation (“trust” and “distrust”), image patterns (all women smiling thus positive, all men frowning thus negative), and causation (“cancer” and “smoke” are both negative and have a cause–effect relationship). Each of these cross-category associations exists because the stimuli themselves have an additional property that allows them to be associated with multiple categories.

Two studies exemplify the effect of cross-category associations on the direction and size of D_{IAT} (for more detailed overviews, see Axt et al., 2021; Greenwald et al., 2021). Steffens and Plewe (2001) kept the category-exemplars (male and female names) constant and varied the gender orientation of the positive attitude-exemplars (e.g., female: beautiful, male: independent). The result was a larger IAT effect when the attitude-exemplars were female-orientated than when the attitude-exemplars were male-orientated. This suggests that cross-category associations between attitude-exemplars and target-categories directly influenced the size of the IAT effect.

A second example of stimulus effects due to cross-category association comes from Bluemke and Friese (2006). They did similar experiments where they manipulated the relationship between the target-categories (East- and West-German nouns and names), the attitude-categories (positive and negative nouns), and the participants. For example, the exemplar “Stasi” was used as a negative exemplar with a cross-category association with former East-Germany. Their experiments showed that IAT effects could be increased if the manipulations favoured the in-group participants (West-Germans), whereas the IAT effects were decreased when the manipulations favoured the outgroup participants (East-Germans). These results suggest that the size of D_{IAT} is affected not only by changes to stimuli but also by changes to the examined participants.

As the discussion above shows, stimuli selection requires careful consideration. This is because stimuli that are unfamiliar or contain cross-category associations have the potential to change the direction and size of IAT effects (D_{IAT}). More importantly, stimulus effects are an interaction between the stimuli characteristics and the participant characteristics. Therefore, by changing either the stimuli or the participants, stimulus effects may be introduced that did not exist before. Even simply 'copy-pasting' an existing and validated IAT to a new research population could be problematic. This is not to say that all previously conducted studies suffer these effects. It could however explain replication issues, contradictory results, or previously found null-results. In conclusion, stimulus (re-)validation is warranted when the stimuli or participant populations change due to the possibility of introducing stimulus unfamiliarity and cross-category associations.

Stimulus validation

Considering the different aspects of measurement validation, it may appear unfeasible to (re-)validate the stimuli for each new study. As a practical solution, Greenwald et al. (2021) proposed two absolute cut-off criteria that can easily be used to determine the suitability of the selected exemplars within the intended research population. They propose that the response data from a small pilot sample should indicate that the exemplars were easily (RTs < 800 ms) and accurately (<10% errors) categorized. Exemplars that do not meet these criteria should be "[…] discarded without further consideration." (p. 7, Greenwald et al., 2021). Greenwald et al. (2021) further suggest that these validation criteria should be applied to data from *pilot* subjects originating from the intended participant population. This allows researchers to account for the stimulus by participant interaction a-priori.

However, researchers also conduct IAT research in situations where the participant population is not known beforehand. For example, 2561 publications[2] utilized data collected by Project Implicit[3]: a website where anyone can take part in IAT research. The demographic characteristics of Project Implicit participants are unknown a-priori and are difficult to predict due to the substantial number of participants each year (e.g., >15,000 participants for the Race-IAT in 2020; see Full sample data section). Researchers who utilize Project Implicit data may thus struggle to validate their stimuli and account for the stimulus by participant interaction from pilot-testing alone. In the current study, we, therefore, applied Greenwald et al. (2021)'s proposed validation criteria as post-hoc validation analyses. Post-hoc analyses undoubtedly draw from the intended participant population, thereby also accounting for the stimulus by participant interaction.

To summarize, in the current research, we applied Greenwald et al. (2021)'s proposed validation criteria as post-hoc analyses to Project Implicit data. In total, we conducted three sets of analyses, which together evaluated the theoretical and practical utility of the criteria as pilot- and post-hoc validation analyses.

Research aims and potential implications

The purpose of our research was to evaluate the proposed validation criteria that exemplars should elicit fast (RT < 800 ms) and accurate (<10% error) participant responses (Greenwald et al., 2021). We applied the proposed criteria across existing IAT data[4] in three sets of analyses. Before conducting these analyses, akin to hypotheses formulation, we thought of general outcome scenarios and their potential implications for existing and subsequent IAT research. We revisit each of the outcome scenarios in the *Discussion*.

In the first analysis, we explored the validity of stimuli across 15 individual IATs. For each IAT, we created 10,000 independent samples of 100 participants and determined stimulus validity within

[2] We extended the search strategy from Greenwald et al. (2020) on August 28th 2023. We conducted an advanced search on the American Psychological Association's PsycNET database (https://psycnet.apa.org/home) for publications including "Implicit Association Test" in the Title, Abstract, Keywords, *OR* Test & Measures *AND* "Project Implicit" in Any Field.

[3] Organization: https://www.projectimplicit.net/; Take-a-Test: https://implicit.harvard.edu/implicit/takeatest.html.

[4] In this study, we used the data from 15 IATs available via Project Implicit (Organization; Take-a-Test). The following IATs are included: Age-IAT, Arab-IAT, Asian-American-IAT, Disability-IAT, Gender-Career-IAT, Gender-Science-IAT, Native-American-IAT, President-IAT, Race-IAT, Religion-IAT, Sexuality-IAT, Skin-Tone-IAT, Transgender-IAT, Weapons-IAT, and Weight-IAT.

each sample. Across the 10,000 re-samples, stimulus validity is likely to vary. The fluctuations between samples provide evidence of the reliability with which one can infer validity from the response data of a random sample of 100 participants.

In an optimal scenario, all stimuli would be deemed valid. However, based on the pilot analyses reported in the Pilot results section, this appeared implausible. A more likely scenario was that at least some IATs contain stimuli that are categorized as invalid. The implications of such findings depend on the assumption of the 'ground-truth'. Assuming the validation criteria are the 'ground-truth', the results would imply that at least some IATs include invalid stimuli. This need not be problematic, as Nosek et al. (2005) clearly show that as little as two stimuli per category can reliably measure IAT effects. A few invalid stimuli may thus simply indicate the need for re-computations of D_{IAT} after the invalid data have been removed. However, the validation criteria themselves have not yet been empirically corroborated. Therefore, assuming that the stimuli have been appropriately selected (i.e., the 'ground-truth'), finding invalid stimuli may also imply a need for optimizing the validation criteria.

In the second analysis, we focused on the context dependency of stimulus validation. The effects of cross-category associations within individual IATs suggest that stimulus validity may also be relative to the context of individual IATs (see Undesirable stimulus effects section). After all, whether a stimulus exhibits cross-category associations with other stimuli depends entirely on the included stimuli. To illustrate, imagine two IATs: a Gender-Career IAT (Men/Women/Career/Family) and a Gender-Criminality IAT (Men/Women/Criminal/Innocent). The name "Jack" as a "Male" stimulus is perfectly inconspicuous in the context of the Gender-Career IAT. At the same time "Jack" has a potential cross-category association in the Gender-Criminality IAT due to Jack the Ripper being a famous male criminal. The 15 IATs included in this study provided a unique opportunity to explore the context dependency of stimulus validity because some stimuli were used across multiple IATs. For example, the Age-IAT and the Skin-Tone-IAT both used the stimuli "Pleasure", "Terrible", and "Evil", allowing for a direct comparison of the validity of stimuli in different IAT contexts. Therefore, in the second analysis, we explored the potential dependency of stimulus validity on IAT context.

Among the possible results of the second analysis are patterns of consistent stimulus (in)validity as well as stimuli that are only (in)valid in some contexts. Stimuli that are consistently (in)valid may imply that some stimuli are especially (un)suited for use in IATs. At the same time, fluctuating (in)validity may imply that the validation criteria were sensitive enough to pick up IAT-dependent stimulus effects (e.g., cross-category associations).

In the third and final analysis, we aimed to determine the effect of stimulus type on stimulus validity. A closer look at the utilized stimuli showed substantial differences with stimuli varying from verbal to visual representations. To illustrate, the Gender-Career IAT exclusively uses nouns to establish the target-categories ("Salary" for the category "Career") but uses names to establish the stereotype-categories ("Michelle" for the category "Female"). This poses the question as to whether the validation criteria proposed by Greenwald et al. (2021) are equally sensitive for different stimulus types. In other words, what is the effect of stimulus type on stimulus validity?

If the results indicate that some stimulus types contain a large number of invalid stimuli, this could be due to issues with response times or accuracy. Invalidation due to response times may imply the need for stimulus-type-specific cut-offs (e.g., images take longer to process than words). However, invalidation due to inaccuracy may imply that some stimulus types are not suited for utilization in IAT paradigms.

Altogether, the three analyses have implications for both the stimuli that have been evaluated as well as the validation criteria themselves. It is important to note that stimulus validity is dependent on the interaction between the stimuli and the participants (see Undesirable stimulus effects section). If we find invalid stimuli in our participant samples, this need not imply that all Project Implicit data are invalid. Each combination of stimuli and participants is unique and should thus be treated accordingly. Our analyses can however point researchers in the direction of where (additional) validation analyses may be needed.

PILOT

We prepared our manuscript with RMarkdown in R (Version 4.2.1; R Core Team, 2020),[5] which had several benefits. First, we reduced research degrees of freedom (Wicherts et al., 2016) by formalizing data preprocessing and exclusion (Exclusion criteria section), the analyses (Pilot analyses section), and to some extent the results (Pilot results section). We also conducted pilot analyses which we used to optimize all code.

A second benefit to preparing and publishing our analyses in R is fostering Open Science. Researchers can easily replicate the analyses on new data sets, use different (model) parameters, and expand with additional analyses or visualizations. All scripts are available via the Open Science Framework (OSF; https://osf.io/dw23y/), which is connected to a GitHub repository. Please see the *workflow* vignette for a practical summary. The full research plan was approved by the Ethics Review Board at the *University of Amsterdam (The Netherlands)*.[6]

Pilot data

We used data provided by Project Implicit[7] – a large-scale data collection project that has collected IAT web responses since 2002 (Greenwald et al., 2003).[8] Visitors to the website must agree to the terms and conditions before they have access to IATs about presidents, body-weight attitudes, race, and more.

The data that Project Implicit provides are freely available via the Open Science Framework (OSF; https://osf.io/y9hiq/) for 16 IATs from 2002 until 2021.[9] The data come in two forms: compressed and raw. The *compressed* data contain one row of information per participant and includes information on demographics (e.g., age, occupation), IAT results (e.g., D_{IAT}), and explicit attitudes (i.e., self-report questions). The compressed data have primarily been used by researchers to determine group-level biases. For example, Charlesworth and Banaji (2019) performed trend analyses of biases from 2007 to 2016, Darling-Hammond et al. (2020) explored the effects of the Corona virus on Asian biases from 2007 to 2020, and Ravary et al. (2019) found that "fat-shaming" incidents predicted spikes in the biases detected with the body-weight IAT.

The *raw* data contain the trial-by-trial information such as IAT parameters (e.g., presented stimulus, category pairing) and response parameters (RT and accuracy). Researchers have used raw response data to, for example, validate new IAT formats (e.g., IAT-recoding free, Rothermund et al., 2009), determine the minimal number of exemplars (Nosek et al., 2005), and determine the effects of random stimulus variation (Wolsiefer et al., 2017).

The raw data and the compressed data can be linked via *session_id*; a unique identifier for each started IAT session. Note that we treat session_ids as if they indicate individual participants. It is technically possible for participants to start multiple sessions per IAT, but the size of the data makes it unlikely that

[5] We, furthermore, used the R-packages *bookdown* (Version 0.31; Xie, 2016), *cowplot* (Version 1.1.1; Wilke, 2020a), *dplyr* (Version 1.0.10; Wickham et al., 2021), *foreign* (Version 0.8.84; R Core Team, 2022), *ggplot* (Wickham, 2016), *ggtext* (Version 0.1.2; Wilke, 2020b), *glue* (Version 1.6.2; Hester, 2020), *here* (Version 1.0.1; Müller, 2020), *knitr* (Version 1.41; Xie, 2022), *lme4* (Version 1.1.31; Bates et al., 2015), *lmerTest* (Version 3.1.3; Kuznetsova et al., 2017), *lubridate* (Version 1.9.0; Grolemund & Wickham, 2011), *papaja* (Version 0.1.1.9001; Aust & Barth, 2020), *readr* (Version 2.1.3; Wickham & Hester, 2020), *rmarkdown* (Version 2.19; Allaire et al., 2022), *Rmisc* (Version 1.5.1; Hope, 2013), *scales* (Version 1.2.1; Wickham & Seidel, 2020), and *tidyverse* (Wickham et al., 2019).

[6] The Ethics Review Board of the University of Amsterdam approved the research plan on 19 March 2021 (ref: 2021-FGW_PSYL-13105).

[7] Jordan Axt, the director of Data and Methodology for Project Implicit, has confirmed on the 16 March 2021 that (1) we did not yet have access to the requested data, and (2) will receive the data after Stage 1 acceptance. The official statement is available via the Open Science Framework (OSF; https://osf.io/dw23y/?view_only=25b62f307a1349e7883549b473). *Stage 2:* Upon requesting access to the full data, we were notified that, since June 2022, all data are now publicly available via the Open Science Framework (https://osf.io/y9hiq/). Please note that these data were made public *after* our preregistration.

[8] Organization: https://www.projectimplicit.net/; Take-a-Test: https://implicit.harvard.edu/implicit/takeatest.html.

[9] Determined at the time of writing (26th January 2022).

a single person contributed a significant number of sessions. We further discuss the issue of repeated measurement in the Exclusion criteria.

For the pilot analyses, we worked with data from the *Gender-Career IAT* (GC-IAT). The compressed data of the GC-IAT were freely available via Project Implicit' OSF page (https://osf.io/abxq7/). We received the raw response data – which was not yet available online – from November and December 2019 via email on 3 September 2020.

We preprocessed the *compressed* data by computing the participants' age at the time of testing, recoding session status to indicate completion, and removing all unnecessary columns (e.g., explicit attitudes). After these preprocessing steps and applying the exclusion criteria (see Compressed data section), the compressed pilot data consisted of 8549 rows of data.

We preprocessed the *raw* data by cleaning block and trial names, determining whether each trial was (in)congruently paired, as well as filtering out data that accidentally belonged to other IATs. After these preparations and applying the exclusion criteria (see Raw data section), the raw pilot data consisted of 1,606,055 rows of data (responses/trials).

Exclusion criteria

The exclusion criteria were based on the available literature and analyses of the pilot data. We differentiate between exclusion criteria that were applied to the compressed data (e.g., demographic criteria) and those that were applied to the raw response data (e.g., extremely long response times). When participants were excluded based on compressed data criteria, their data were also removed from the raw data. Figures 1 and 2 show the impact of the exclusion criteria when applied to the pilot and full data, respectively.

Compressed data

The compressed data contain information on summary statistics of the IAT, demographics, and answers to explicit questions (removed during data preprocessing). Participants who did not complete the full IAT were removed from further analyses based on missing D_{IAT} and/or an incomplete *session_status*.

In the new guidelines, Greenwald et al. (2021) state that the IAT retains its' measurement properties with repeated measurement, although evidence indicates more polarized results at the first IAT measurement. We thus opted to refrain from exclusion based on prior IAT experience.

Greenwald et al. (2003) showed the statistical benefits of removing participants who made too many fast responses. Participants who responded faster than 300 ms on more than 10% of the trials were excluded from subsequent analyses (Step 5 in appendix B in Greenwald et al., 2021).

The criteria discussed above focused on aspects inherent to the IAT paradigm, but two additional criteria – based on demographic information – warrant consideration. First, to determine the validity of IAT stimuli, we need to ensure that participants have the highest common ground with regard to concept meaning. We cannot control for this fully but restricting the participant sample to participants who live and grew up in the same country is the best approximation possible for the current data set. We thus excluded participants that did not reside in, nor have citizenship of the United States of America (USA). We further excluded participants who opted not to provide citizenship and/or residency information.

Second, when participants enter the Project Implicit website, they are asked to provide informed consent. Even though participants may have agreed to the terms and conditions, US States enforce age-limits for the ability to provide consent (ranging from 16 to 18; https://www.ageofconsent.net/states). We excluded participants who self-reported to be younger than 18 years old at the time of IAT completion. Greenwald et al. (2021) further proposed that the validity of exemplars may be derived from examination of the data of "young adult subjects" (p. 7). We defined "young adult subjects" as participants between 18 and 25 years old, excluding all participants who self-reported to be older than 25.

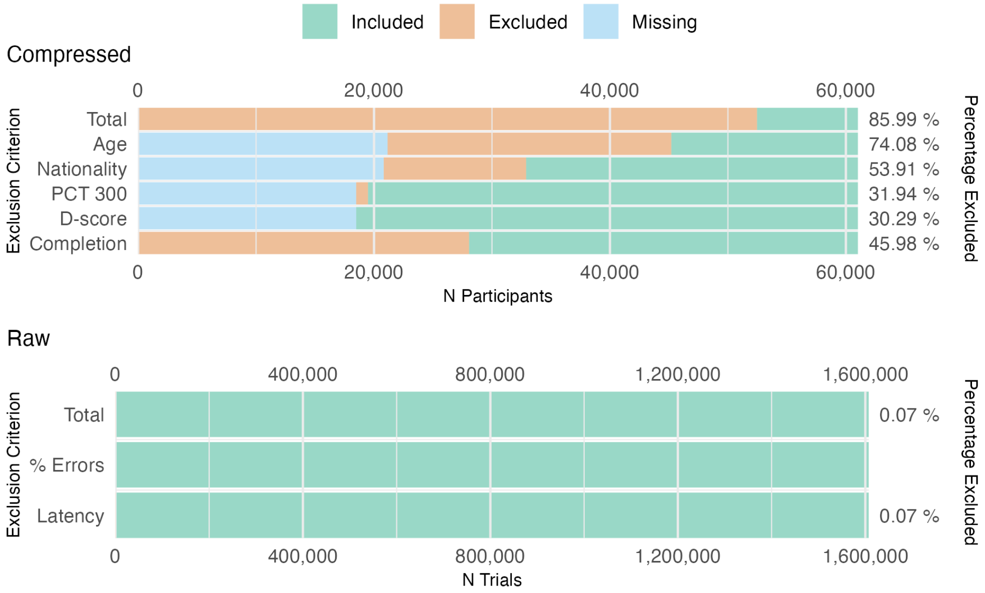


FIGURE 1 Summary of excluded pilot data. (a) The effects of the exclusion criteria on the *compressed* data. *Completion*: participants were excluded if they failed to complete the full IAT. *D-score*: for various reasons, among which technical error, IAT D-scores (i.e., D_{IAT}) may have been missing. *PCT 300*: participants were excluded if the percentage of responses faster than 300 ms was higher than 10%. *Nationality*: participants were excluded if they did not reside in, nor have citizenship of, the United States of America. *Age*: participants were excluded if they were younger than 18 or older than 25. *Total*: the unique number of participants excluded based on missing data and/or explicit exclusion. (b) The effects of the exclusion criteria in the *raw* pilot data. *Latency*: trials were excluded if they exceeded 10,000 ms or were 0 ms. *% Errors*: trials were excluded if the participant answered more than 50% of their trials incorrectly. *Total*: the total number of trials excluded based on the prior criteria. *Note*: the data of participants, which were excluded based on the *compressed* data, were excluded from the *raw* data *prior* to applying the exclusion criteria. After mutual exclusion, 8549 participants were eligible for analyses.

Raw data

The raw data include trial-by-trial information such as response times and errors. Previously, slow and/or fast response times were deleted or recoded. In line with guidelines set forth by Greenwald et al. (2003, 2021), we applied an upper limit excluding trials where the response time is larger than 10,000 ms. We did not recode responses that could be considered as "too fast" (e.g., <400 ms) because systematic too fast responses may offer relevant information about the variation in average response times. We did, however, exclude responses with a response time of zero milliseconds as these were likely the cause of technical malfunction. Note that the trials are removed, but the participants themselves were not excluded.

Contrary to suggested guidelines, in the present research, we excluded participants based on error percentages. Greenwald et al. (2003) showed that it was unnecessary, but not incorrect, to exclude participants with high error rates (Study 3). We opted to exclude participants who performed below chance (<50% correct trials across all blocks), because the low success rates may indicate that participants did not understand task demands and should thus not be considered when validating task content.

Pilot analyses

We determined stimulus validity by applying the validation criteria (Validation criteria section) to the response data of 10,000 samples of 100 participants (Sample size and Bootstrapping sections). These re-samples, also known as *m* out of *n* bootstraps, together illustrated how reliably one can infer stimulus

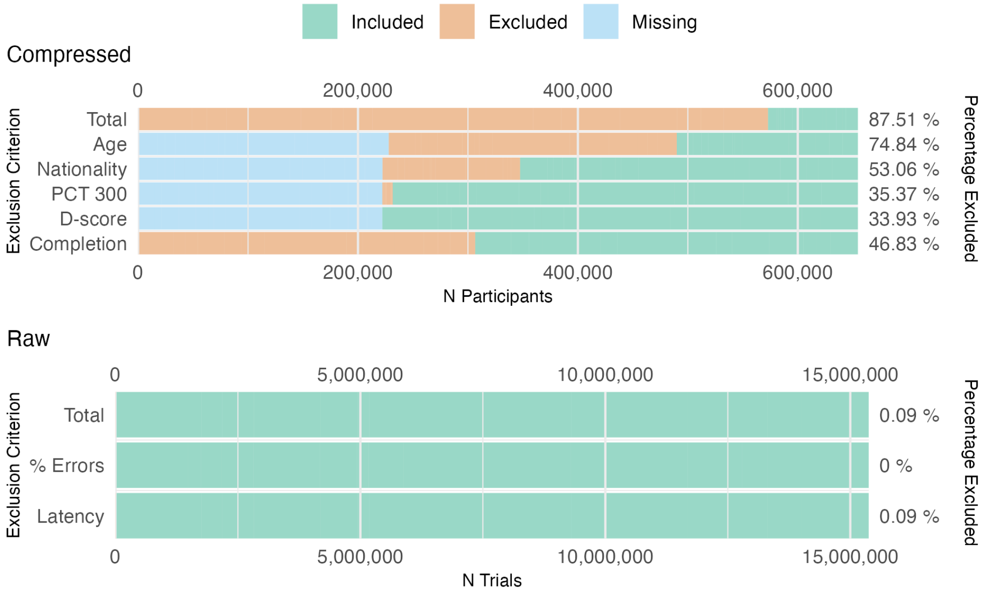


FIGURE 2 Summary of the excluded full data. (a) The effects of the exclusion criteria on the *compressed* data. *Completion*: participants were excluded if they failed to complete the full IAT. *D-score*: for various reasons, among which technical error, IAT D-scores (i.e., D_{IAT}) may have been missing. *PCT 300*: participants were excluded if the percentage of responses faster than 300 ms was higher than 10%. *Nationality*: participants were excluded if they did not reside in, nor have citizenship of, the United States of America. *Age*: participants were excluded if they were younger than 18 or older than 25. *Total*: the unique number of participants excluded based on missing data and/or explicit exclusion. (b) The effects of the exclusion criteria in the *raw* data. *Latency*: trials were excluded if they exceeded 10,000 ms or were 0 ms. *% Errors*: trials were excluded if the participant answered more than 50% of their trials incorrectly. *Total*: the total number of trials excluded based on the prior criteria. *Note*: the data of participants, which were excluded based on the *compressed* data, were excluded from the *raw* data *prior* to applying the exclusion criteria. After exclusion, 81,789 participants were included in total (N = 1269–18,178, M = 5453, SD = 4260).

validity from a random sample of 100 participants (Reliability section). The bootstrap procedure and results visualization (Pilot results section) were formalized for the pilot analyses of the Gender-Career IAT (November–December, 2019) during Stage 1.

Validation criteria

We implemented the validation criteria proposed by Greenwald et al. (2021):

> A judgement as to whether specific exemplars are easy enough to classify can be based on examination of data obtained from pilot subjects. The useful data will come from Blocks 1 and 2 of the standard procedure (see Appendix A). Pilot subjects should be able to categorize all stimuli in these two blocks rapidly (average latency in the range of 600–800 ms for most young adult subjects) and with low error rates (less than 10%).
>
> (Greenwald et al., 2021, secs. 1–A8, p. 7)

According to Greenwald et al. (2021), stimulus validity is thus inferred from two parameters computed from the responses of a (pilot) sample of participants. First, within a sample of participants, the average response time should be faster than 800 milliseconds. Second, those participants should categorize the stimulus incorrectly in less than 10% of the trials. These criteria – independent of the sample size – thus result in a dichotomous decision (yes/no) of stimulus validity: a stimulus is deemed valid if *both* the average response time and error rates are below the specified thresholds.

Sample size

Greenwald et al. (2021) proposed that the validity of exemplars (i.e., stimuli) may be derived from "[…] on examination of data obtained from pilot subjects," while at the same time, they stated: "Subjects for pilot testing should come from the intended research subject population" (p. 7). The latter is per definition the case in post-hoc analyses of experimental data.

Validating stimuli from experimental data ensures that the validity is always inferred from the research population of interest. This is especially relevant when we consider that (un)familiarity of stimuli and cross-category associations – as discussed in Undesirable stimulus effects section – are likely to differ among populations. Garimella et al. (2017), for example, showed that free-associations differed depending on gender (Male/Female) and location (USA/India). The cue "bath" was associated with "water" for Males irrespective of their location, but with "bubbles" for US Females and "soap" for Indian Females. Validation of the stimuli from experimental data may thus provide researchers with evidence of unfamiliarity or cross-category associations that are specific to their research population.

To determine a feasible experimental sample size, we looked at the sample sizes of studies included in some of the published meta-analyses. Greenwald et al. (2009) included 184 independent samples with an average of 81 participants (*SD* = 141.53, Min = 9, Max = 1386). Babchishin et al. (2013) included 12 studies with an average of 66 participants (*SD* = 24.39, Min = 38, Max = 113). Finally, Oswald et al. (2013) included 97 independent samples with an average of 65 participants (*SD* = 113.66, Min = 12, Max = 1057). To err on the side of caution, the following analyses relied on a sample size that is somewhat higher than the averages of the reported meta-analyses: 100 participants. In doing so, we increased the chances of finding true effects within a sample (i.e., power) while staying within reach of what is feasible for (future) experimental studies.

Bootstrapping

Although 100 participants are a feasible sample for experimental studies, the data provided by Project Implicit are much more extensive (e.g., $N_{\text{pilot}} = 8549$). The large number of participants provided the opportunity to simulate the results of conducting 10,000 'experiments' (i.e., samples) per IAT with a sample size of 100 participants each. In other words, we conducted 10,000 *m* out of *n* bootstraps (Bickel et al., 2012), where *m* is the number of sampled participants (100) and *n* is the number of available participants. Note that the total number of available participants (*n*) differed across IATs.

For each sample of 100 participants, we determined the average response time and error rate per stimulus; classifying stimuli as valid if the average response time was less than 800 milliseconds and the error rate was less than 10%. In total, we thus had 10,000 classifications of validity per stimulus per IAT. The percentage of 10,000 samples in which a stimulus was classed as valid is denoted as the *percentage_valid*.

Reliability

The percentage_valid indicates how reliably one can infer stimulus (in)validity from a sample of 100 participants. We defined stimuli as *reliably valid* if the stimulus was valid in 95% or more of the 10,000 samples. Similarly, a stimulus was classed as *reliably invalid* if the stimulus was valid in 5% or less of the samples. If stimuli were classed as valid in 6%–94% of the samples, we classed them as *unreliable* to indicate that we are less than 5% certain that a new sample of 100 participants would yield the same (in) validity judgement.

We determined whether stimuli were reliably valid for the response time and error rate criteria separately. This provided a better insight into which of the two criteria has the biggest impact on judgements of stimulus validity. However, Greenwald et al. (2021) clearly describe that valid stimuli are rapidly (response time) *and* accurately (error rate) categorized. We thus also computed a final validity judgement where a stimulus is *valid* if both the response time and error rate were reliably valid but *invalid* if either of the criteria was reliably invalid. Any combination of valid-unreliable or unreliable-unreliable resulted in a final judgement of *unreliable* as we could not be sure about the validity of both criteria.

Pilot results

We conducted pilot analyses with the data of the GC-IAT to formalize the analyses and reports of the results for individual IATs. The Gender-Career-IAT included 8549 participants who provided 1,606,055 responses overall and 341,699 responses in Blocks 1 and 2. The included participants were on average 20.67 years old (*SD* = 2.43; 95% CI = 20.72, 20.62).

Figure 3 shows the distribution of average response times and error rates across 10,000 samples of 100 participants of the pilot GC-IAT. Based on average RT, eight of the stimuli used in the Gender-Career-IAT were reliably valid (33.33%), eight reliably invalid (33.33%), and eight could not be reliably

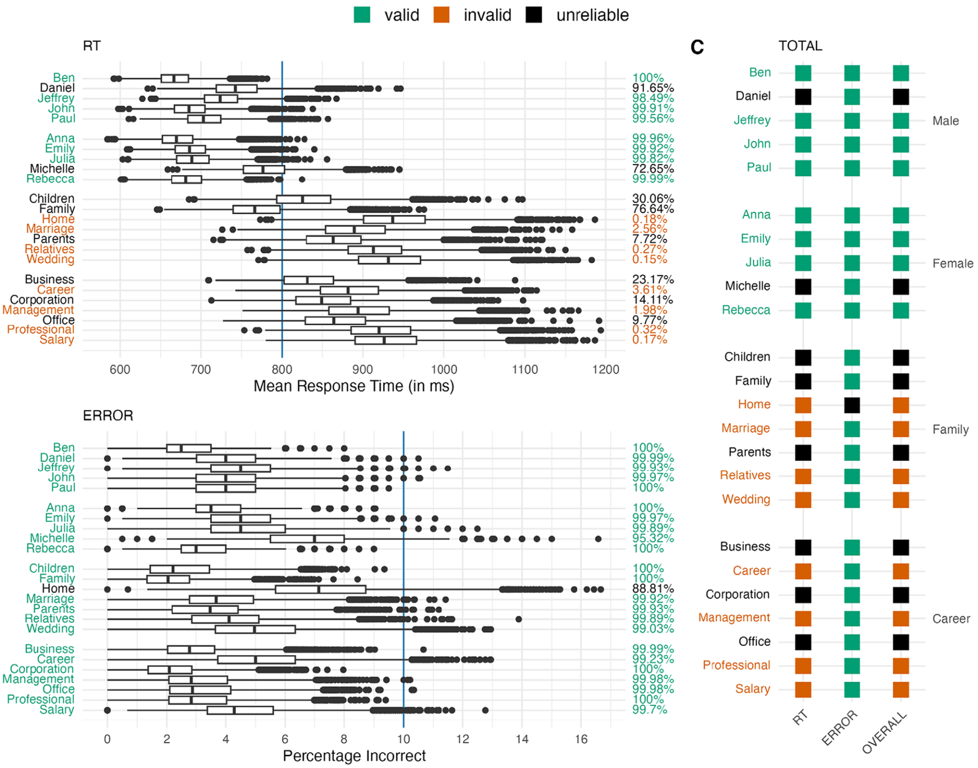


FIGURE 3 The validity estimates per stimulus of the Gender-Career IAT (November–December, 2019). [RT] The distribution of *average response time* across 10,000 samples of 100 participants. Within each sample, a stimulus (left *y*-axis) is judged as valid if the average response time is lower than 800 milliseconds (vertical blue line). A stimulus is classed as *reliably* valid (green) if 95% or more of the samples resulted in a valid judgement (right *y*-axis). Stimuli are considered reliably invalid (red) if a stimulus is classed as valid in 5% or less of the samples. Stimuli that were valid in 6% to 94% of the samples were classed as unreliable (black). [ERROR] The distribution of the *percentage of errors* across 10,000 samples of 100 participants. Within each sample, a stimulus (left *y*-axis) is judged as valid if less than 10 per cent of the trials were answered incorrectly (vertical blue line). [TOTAL] An overview of the validity judgements per exemplar based on average response times (RT) and percentage of errors (ERROR). A stimulus was classed as valid (TOTAL) if both criteria were reliably valid, but as invalid if either criterion was reliably invalid. If the criteria were both unreliable, a stimulus was also classed as unreliable.

estimated (33.33%). With respect to the percentage of errors, 23 of the stimuli used in the Gender-Career-IAT were reliably valid (95.83%), 0 reliably invalid (0%), and 1 could not be reliably estimated (4.17%). Taken together, based on RT *and* error percentages, eight of the stimuli used in the Gender-Career-IAT were reliably valid (33.33%), eight reliably invalid (33.33%), and eight could not be reliably estimated (33.33%).

The pilot analyses of the GC-IAT showed interesting patterns, which served as input for the preregistered analyses (Pilot analyses section). First, Figure 3 shows that the RT and ERROR criteria differently conclude stimulus (in)validity. Based on the RT criterion, Male/Female names are predominantly considered reliably valid (green), whereas Career/Family nouns are predominantly considered reliably invalid (red). This distinction is not evident for the ERROR criterion where all stimuli but 'Home' are reliably valid. The distinction is again evident for the TOTAL criterion, where the RT and ERROR validity are combined. Because both criteria must conclude stimulus validity, the criterion with the highest levels of stimulus invalidity – in this case the RT criterion – dominates the final validity verdict (TOTAL). We report on the RT and ERROR criterion separately rather than solely on the TOTAL to show when conclusions differ.

The pilot results showed three stimuli, which specifically stand out. 'Michelle' and 'Daniel' were the only two unreliable stimuli based on the RT criterion. The third stimulus, 'Home', was the only not reliably valid stimulus based on the ERROR criterion. The fact that individual stimuli could be problematic prompted us to explore whether stimulus (in)validity is consistent across IAT contexts (see Contextual differences section).

Finally, Figure 3 shows that the average response times of the Career/Family stimuli (nouns) was higher than in the Male/Female stimuli (names). These differences were so substantial that names were deemed reliably valid much more than the nouns. The differences between the types of stimuli prompted us to preregister the third analysis where we explored the relationship between stimulus validity and stimulus type.

FULL SAMPLE

The full analyses extent the pilot in two aspects. First, we applied the pilot analyses to the full data set of 15 IATs (November, 2020). The bootstrapped validity estimates were used for two additional – preregistered – analyses. We explored the extent to which stimulus validity differs across IAT contexts (Contextual differences section). Then, we explored the effect of stimulus type on stimulus validity (Stimulus type and validity section). For these two analyses, only the analysis plans were preregistered.

The pilot code was not capable of (1) downloading data directly from the Open Science Framework, (2) dealing with larger datasets, and (3) visualizing the final results in a comprehensive manner. We therefore updated the code while keeping track of all changes with git version control. The changes were also documented in an explanation of changes, allowing anyone to track and verify (the need for) the changes we made between Stage 1 (pilot) and Stage 2 (full sample).

Full sample data

We utilized the data from 15 IATs: the Age-IAT, Arab-IAT, Asian-American-IAT, Disability-IAT, Gender-Career-IAT, Gender-Science-IAT, Native-American-IAT, President-IAT, Race-IAT, Religion-IAT, Sexuality-IAT, Skin-Tone-IAT, Transgender-IAT, Weapons-IAT, and Weight-IAT. We specifically used the data from November 2020, which were downloaded directly from the relevant OSF repositories (see Online Appendix datasets/OSF-urls.).

After downloading the data, we preprocessed the *compressed* data by (1) filtering the data for the *months_of_interest* (Nov), (2) recoding session status to indicate completion, (3) computing the participants' age at the time of testing, and (4) removing all unnecessary columns (e.g., explicit attitudes). Note that the first preprocessing step – filtering for months_of_interest – was not explicitly preregistered. We added this filter during Stage 2 to reduce the amount of time spent preprocessing the data. A requirement that became critical only *after* trying to process the data of the more popular IATs. For example, the Race-IAT contains 2.27 GB of data for 2020, whereas the Gender-Career IAT contains only 267 MB. Although an explicit filter was not part of the preregistered analysis code, we did in fact indirectly filter for months_of_interest. As we only received the raw data from November/December 2019 and excluding vice versa meant that we excluded any participants from which we did not have raw data. Explicitly filtering for months_of_interest therefore only reduces the amount of data being preprocessed, but not the actual data used for the validation analyses.

Next, we preprocessed the *raw* data by (1) retaining only the data from session_ids included in the preprocessed compressed data, (2) cleaning block and trial names, (3) determining whether each trial was (in)congruently paired, and (4) filtering out data that accidentally belonged to other IATs. The first preprocessing step – filtering for session_ids – was not preregistered but was added in Stage 2 to reduce preprocessing time. The need for an additional preprocessing step again arose from large datasets. We therefore opted to preprocess only the raw data of participants which completed the IAT during the months_of_interest. Critically, the raw data, unlike the compressed data, do not contain information on when the data were collected (i.e., a date). Therefore, filtering by session_id of the compressed data served as a proxy for applying a months_of_interest filter on the raw data.

After these preparations and applying the exclusion criteria (see Raw data section), the Race-IAT had the most eligible participants ($N = 18{,}178$) and the Native-American-IAT the least ($N = 1269$; Mean = 5453; $SD = 4260$).

Full sample analyses

We first computed the bootstrapped validity estimates for each of the 15 IATs (Bootstrapping section). We then explored the extent to which stimulus validity differs across IAT contexts (Contextual differences section). Finally, in the third analysis, we explored the effect of stimulus type on stimulus validity (Stimulus type and validity section).

Contextual differences

The 15 IATs included in this study contained a total of 395 unique stimuli, out of which 64 stimuli were used in nine/ten IATs. These stimuli allowed us to determine whether stimulus (in)validity differed across IAT contexts. Derivatives of the same stem (e.g., “Friend” and “Friendship”) were included separately to prevent variations due to unknown lexical-syntactic properties. It was technically possible for images ($N = 236$) to be included in these analyses. However, solely based on *trial_names* in the *Raw* data (e.g., ‘abled1.jpg’), no images appeared to be reused across IATs.

Contextual differences are inferred from variability in the percentage_valid across the nine/ten IATs in which the stimulus was used. Larger variability indicates that the percentage of 10,000 samples in which a stimulus was classed as valid depends on the IAT in which the stimulus was presented. We thus infer between-IAT variability from visual inspections of raw data and 95% Confidence Intervals of the Mean.

Stimulus type and validity

In the final analysis, we explored whether the validation criteria proposed by Greenwald et al. (2021) were equally sensitive for different stimulus types. We first determined which of the 395 stimuli

TABLE 1 Stimulus types across 395 stimuli from 15 IATs.

Stimulus type	*N*	Examples	Mixed-effects model
Image	236	tone0031a.jpg, nixon1.jpg, of3.jpg, recent15.jpg, jefferson2.jpg…	Included
Adjective	69	Horrible, Scorn, Disaster, Poison, Friendship…	Included
Noun	57	Boy, Judaism, Husband, Muslim, Humanities…	Included
Name	29	Jeffrey, Habib, Yousef, Hakim, Michelle…	Included
Multi-Word	4	Gay People, Gay Men, Gay Women, Straight People	Excluded

Note: The Examples column displayed five randomly selected stimuli. The final column indicates whether the stimulus types were included as part of the mixed-effects models.

were 'Images' ($N=236$) based on pattern recognition (.jpg & .png). We then manually assigned *stimulus_type* (Image, Adjective, Noun, Name, and Multi-Word) to the remaining stimuli ($N=159$; see Table 1).

We planned to fit a mixed-effects regression model analysis with *percentage_valid* as the dependent variable and *stimulus_type* as a fixed-effect. In addition, as stimulus types are nested within IATs, we also wanted to account for both between-IAT and within-IAT effects. We wanted to account for between-IAT effects by including *IAT* as a random-effect: each *IAT* was fitted with a unique *percentage_valid* intercept. To account for the possibility that stimulus types nested within IATs exerted an effect on validity, we wanted to include *stimulus_type* as a random slope: the relationship between *percentage_valid* and *stimulus_type* can differ between IATs:

$$percantage\_valid \sim 1 + stimulus\_type + \left(1 + stimulus\_type | IAT\right)$$

We fitted the mixed-effects model with the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages with a BOBYQA optimizer (Powell, 2009) and a maximum of 200,000 iterations (Miller, 2018). Unfortunately, the model did not converge due to issues with (near) singularity. We simplified the model by removing the random-effects of *stimulus_type*, because each IAT only contains one or two stimulus types. Removing the random-effects therefore offers the greatest chances of successfully fitting a mixed-effects model, while still accounting for the fact that stimulus validity may differ across IATs.

We thus fitted the following model for the *percentage_valid* for the response time (RT), error rate (ERROR), and overall (TOTAL) validity criteria:

$$percentage\_valid \sim 1 + stimulus\_type + (1|IAT)$$

Full sample results

Stimulus validity

In a direct replication of the pilot analyses, we first determined stimulus validity within each of the 15 included IATs. The results of individual IATs are available in the Online Appendix. The pilot results of the Gender-Career IAT (2019) and the preregistered analyses (2020) were approximately equal. The conclusions of stimulus validity were identical, even though there were minor variations in the bootstrapped percentages of samples in which the stimulus was valid (see the Online Appendix). Figure 4 depicts stimulus validity across all stimuli ($N_{\text{unique}}=395$) included in the 15 IATs. For illustrative purposes, we have also included the results from the most popular IAT: the Race-IAT (Figure 5). The Race-IAT included 18,178 participants who provided 3,413,873 responses overall and 726,134 responses in Blocks 1 and 2. The included participants were on average 20.84 years old ($SD=2.46$; 95% CI = 20.88, 20.81).

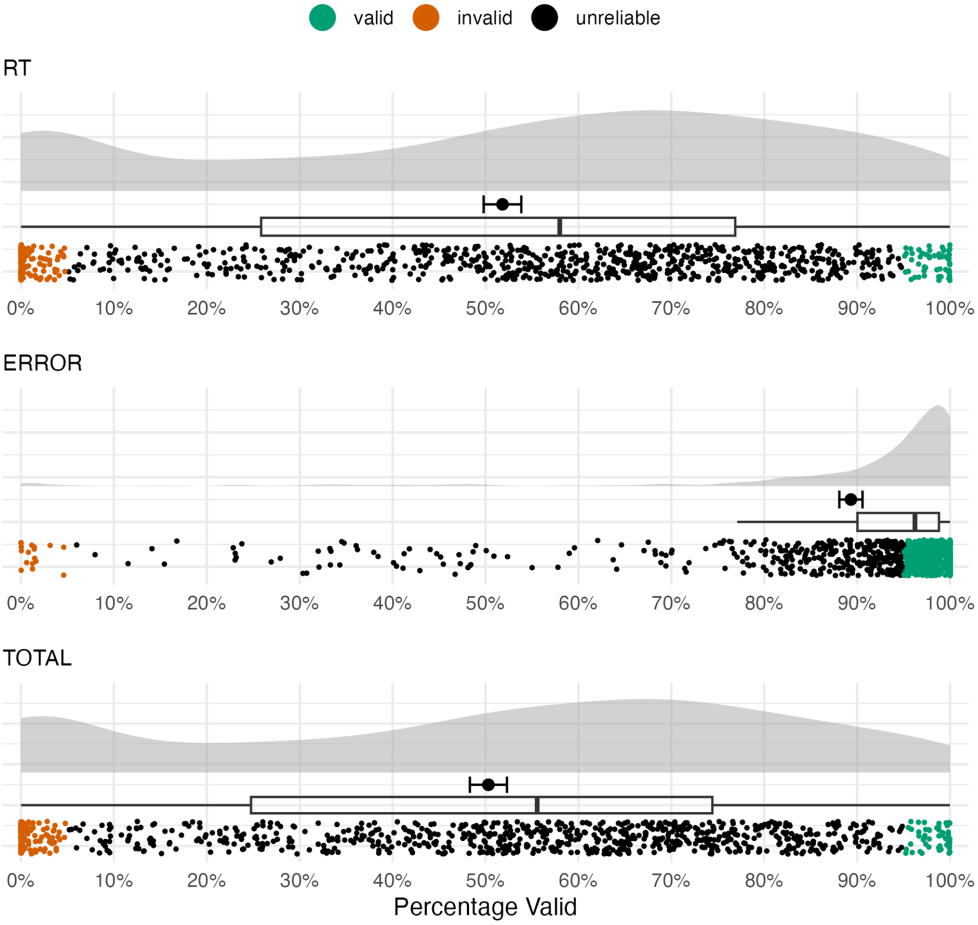
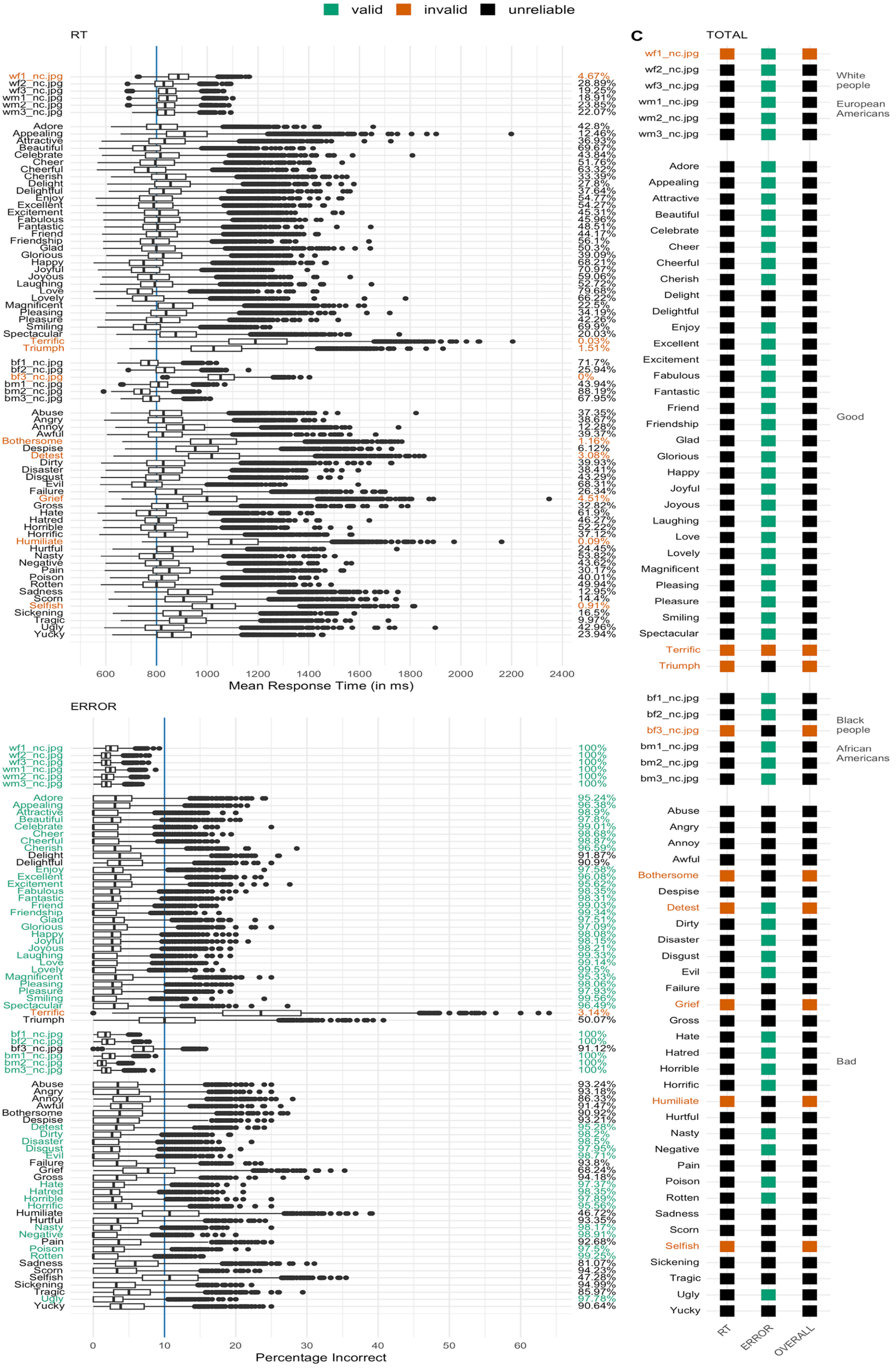


FIGURE 4 Distributions of the percentage valid (x-axis) of all stimuli (dots; $N_{\text{total}} = 923$). The percentage valid represents the percentage of 10,000 samples of 100 participants in which a stimulus was valid based on the respective validation criteria. Each subplot from top-to-bottom shows the (1) global distribution, (2) mean and 95% Confidence Interval, (3) a standard boxplot without outliers, and (4) the raw data (individual stimuli). [RT] Stimuli were classed valid if the average response time of the participant sample was lower than 800 ms. [ERROR] Stimuli were classed valid if the error rate of the participant sample was lower than 10%. [TOTAL] Stimuli were valid if both the RT and ERROR criteria were valid, invalid if either criterion was invalid, and unreliable if either was unreliable.

Across 15 IATs, we determined stimulus validity for 923 stimuli-IAT combinations ($N_{\text{unique}} = 395$). Based on average RT, 58 of the stimuli used in all IATs were reliably valid (6.28%), 142 reliably invalid (15.38%), and 723 could not be reliably estimated (78.33%). With respect to the percentage of errors, 530 of the stimuli used in all IATs were reliably valid (57.42%), 17 reliably invalid (1.84%), and 376 could not be reliably estimated (40.74%). Taken together, based on RT *and* error percentages, 54 of the stimuli used in all IATs were reliably valid (5.85%), 145 reliably invalid (15.71%), and 724 could not be reliably estimated (78.44%). The validity estimates (percentage_valid) were used in two subsequent preregistered analyses.

Based on average RT, 0 of the stimuli used in the Race-IAT were reliably valid (0%), 9 reliably invalid (11.84%), and 67 could not be reliably estimated (88.16%). With respect to the percentage of errors, 53 of the stimuli used in the Race-IAT were reliably valid (69.74%), 1 reliably invalid (1.32%), and 22 could not be reliably estimated (28.95%). Taken together, based on RT *and* error percentages, 0 of the stimuli used in the Race-IAT were reliably valid (0%), 9 reliably invalid (11.84%), and 67 could not be reliably estimated (88.16%). The pattern of stimulus (in)validity in the Race-IAT is consistent with that of the



other IATs (see Online Appendix). We consistently see a pattern where the overall stimulus validity (TOTAL) mimics the RT rather than the ERROR criterion. This is because a stimulus can only be considered valid if *both* criterion are met, whereas a stimulus is already invalid if *one* criterion concludes

FIGURE 5 The validity estimates per stimulus of the Race-IAT (November, 2020). [RT] The distribution of *average response time* across 10,000 samples of 100 participants. Within each sample, a stimulus (left *y*-axis) is judged as valid if the average response time is lower than 800 milliseconds (vertical blue line). A stimulus is classed as *reliably* valid (green) if 95% or more of the samples resulted in a valid judgement (right *y*-axis). Stimuli are considered reliably invalid (red) if a stimulus is classed as valid in 5% or less of the samples. Stimuli that were valid in 6%–94% of the samples were classed as unreliable (black). [ERROR] The distribution of the *percentage of errors* across 10,000 samples of 100 participants. Within each sample, a stimulus (left *y*-axis) is judged as valid if less than 10 per cent of the trials were answered incorrectly (vertical blue line). [TOTAL] An overview of the validity judgements per exemplar based on average response times (RT) and percentage of errors (ERROR). A stimulus was classed as valid (TOTAL) if both criteria were reliably valid, but as invalid if either criterion was reliably invalid. If the criteria were both unreliable, a stimulus was also classed as unreliable.

stimulus invalidity. As the RT criterion concludes more stimulus invalidity, the final verdict therefore closely resembles that of the RT criterion.

The results from the Race-IATs illustrates another pattern which we consistently see across the 15 included IATs: between-sample variance. The width of the boxplots indicates that the average response time and error rate vary greatly between the 10,000 samples. As a consequence, most stimuli could not be reliably (in)validated, indicating that we were frequently less than 5% sure that a new sample of 100 US participants would yield the same (in)validity conclusion.

Contextual differences

In the second analysis, we explored whether stimulus (in)validity differed across IAT contexts. A total of 64 stimuli were used in nine/ten different IATs. Between-IAT variance in stimulus validity was used to infer contextual differences. Figure 6 illustrates the differences in stimulus validity across IAT contexts.

Figure 6 shows that stimulus validity varies between IATs. This is evident from the between-IAT variation displayed as 95% Confidence Intervals of the Mean. Only some items demonstrate context independence: the 95% CIs are extremely small. This context independence manifests on the extremes of the percentage_valid scales (e.g., “Smiling”, “Terrific”). Stimuli are either consistently invalid or consistently valid.

Contextual differences are however evident when stimulus validity could not be reliably estimated. Each stimulus in the unreliable region displays large between-IAT variability (e.g., “Tragic”). Within those stimuli, it is evident that some IATs (points) score considerably lower/higher than the other IATs on the percentage of samples in which the stimulus was valid. The under-performing IAT is generally speaking the Race-IAT when the RT criterion is concerned, but the Arab-IAT for the ERROR criterion. The over-performing IAT with respect to RT is most often the President-IAT, but with the ERROR criterion, both the Race- and Transgender-IAT outperform compared to the other IATs.

Stimulus types

In the pilot analyses of the Gender-Career IAT (2019), we saw differences in validity between two stimulus types (nouns; names). The third analysis therefore explored the effect of stimulus type on stimulus validity. We fitted a mixed-effects model with random effects for IAT (controlling for contextual differences) and fixed-effects stimulus types. We included the stimulus types “Images” (intercept), “Adjectives,” “Nouns,” and “Names” (proper nouns; see Stimulus type and validity section). The fixed-effects and the uncontrolled raw data are depicted in Figure 7.

All stimulus types, bar the nouns in the ERROR model, were significant predictors of percentage_valid when corrected for IAT context. The fixed effects show that images, on average, are valid in more of the 10,000 samples than other stimulus types. This is the case for both the RT and ERROR criterion, although the effect of stimulus type is more pronounced for the RT criterion. The names in

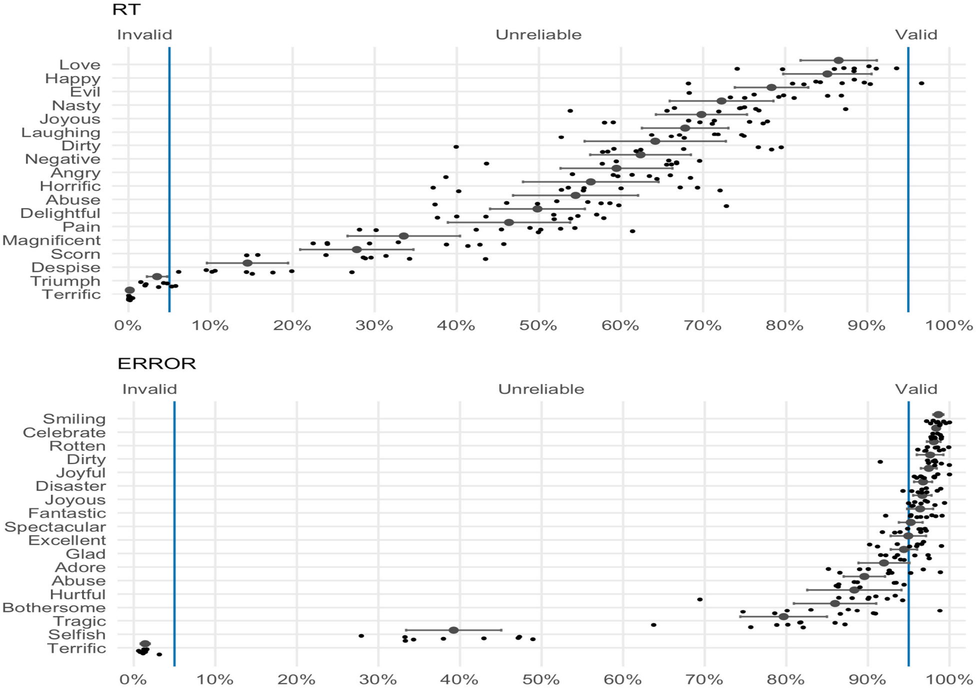


FIGURE 6 The percentage of samples in which a stimulus was valid for the nine/ten different IATs in which the stimulus occurred. The percentage valid represents the percentage of 10,000 samples of 100 participants in which a stimulus was valid based on the respective validation criteria. From top-to-bottom, each plot shows (1) the 95% Confidence Interval of the Mean and (2) the raw data (individual IATs). This plot includes a subset of stimuli (every 4th stimulus) which showcases the general patterns while still keeping the plot legible. A full-sized zoomable plot is available in the Online Appendix.

the ERROR criterion are also a significant predictor of stimulus validity, yet are considerably more error prone than other stimulus types.

When we explore the raw data, it becomes evident that there is large variability of stimulus validity within stimulus types. To some extent, this variability was already evident in the analyses of contextual differences – as the reused stimuli were all adjectives. Variability within stimulus types is also evident in the individual IAT results (see the Online Appendix). For example, in the President-IAT, the images of the recent presidents (Trump; Obama) were valid, while images of older presents were unreliable. This illustrates that there is both within- and between-stimulus type variability.

DISCUSSION

The *Implicit Association Test* (IAT, Greenwald et al., 1998) is a popular measurement of implicit attitudes and (stereotypical) biases. After two decades of IAT research, Greenwald et al. (2021) published "Best Practices in IAT Research": practical guidelines for designing and administering IATs. Among those guidelines were the recommendation to include only those stimuli which a small representative pilot sample is able to classify with speed (RT < 800 ms) and accuracy (< 10% errors; p. 7). We explored the theoretical and practical utility of these criteria for stimulus validation in three sets of preregistered analyses.

We first determined stimulus validity in 15 IATs currently available on Project Implicit. The results show that each of the evaluated IATs contained multiple invalid stimuli according to the validation criteria proposed by Greenwald et al. (2021). Overall, only 5.85% of the stimuli were reliably valid (*N* = 54), while 15.71% of the stimuli were reliably invalid (*N* = 145). Most stimuli, however, could not be reliably

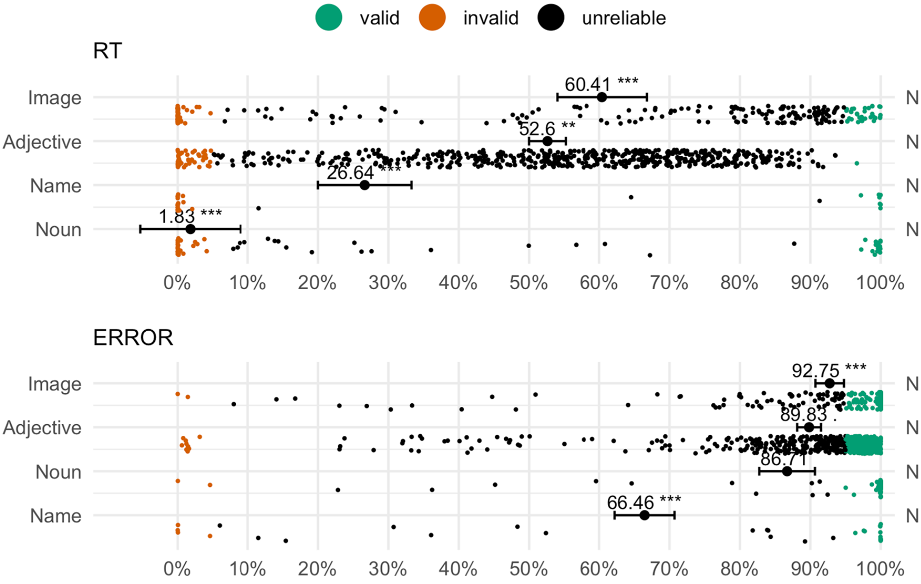


FIGURE 7 The fixed-effects of two mixed-effects models predicting the percentage_valid (*x*-axis) from stimulus type (*y*-axis) corrected for IAT context (not visualized). Each subplot shows the fixed-effect estimate (point) with 95% Confidence Interval (error bars), and the uncontrolled raw data (individual stimuli). ***$p<.001$; **$p<.01$, *$p<.01$, . $p<.1$.

(in)validated. For 78.44% of the stimuli, we could not predict with certainty that a new sample of 100 participants would yield similar (in)validity estimates.

In the second analysis, we found that stimulus validity differed across IAT contexts. Low between-IAT variance in stimulus validity was evident for both invalid (e.g., "Terrific") and valid (e.g., "Smiling") stimuli. Contextual differences (i.e., high between-IAT variance) occurred when validity could not be reliably estimated. Explorations of the between-IAT variance showed that specific IATs were consistent 'outliers': the percentage of samples in which a stimulus was valid differed considerably for one IAT compared to the other nine/ten. With respect to the RT criterion, the Race-IAT consistently scored lower on the percentage of samples in which a stimulus was valid. For the ERROR criterion, this was the Arab-IAT. The occurrence of (consistent) outlying IATs indicates that stimulus validity was at least to some extent different across IAT contexts.

In the final analysis, we explored whether stimulus type (images; nouns; names; adjectives) predicted stimulus validity. Two mixed-effects models – where we controlled for IAT context – showed that stimulus types were significant predictors of stimulus validity. The raw uncontrolled data in addition showed that stimulus validity varied greatly within stimulus types. For example, images were on average most valid, but the Native-American IAT contained multiple images, which were consistently categorized too slow and incorrectly (i.e., invalid).

Greenwald et al. (2021) stated: "Exemplars that just one of a small group of pilot subjects finds difficult to classify are safely discarded without further consideration" (A8). Our analyses therefore indicate that stimuli should be removed from all 15 evaluated IATs. We hypothesized in the Research aims and potential implications section that removing stimuli post-hoc need not be problematic because Nosek et al. (2005) showed that single-stimulus IATs were still valid measurements. However, in some IATs, entire categories failed to meet the validation criteria (e.g., "Other People" in the Arab-IAT), which would prevent computation of D_{IAT} (see Appendix A). In addition, Greenwald et al. (2021) recommend that each category is represented by at least three exemplars (table 1, B6). This would not be possible if only the valid stimuli are retained. Strictly applying the validation criteria would thus require the 15 evaluated IATs to be completely redesigned.

It is important to note however that the bulk of items could not be reliably validated (78.44%). Rather than discarding all these unreliable stimuli, we should consider when unreliable estimates occur and how we can account for them. The three analyses together indicate that unreliable estimates are caused by three factors: between-sample variance, between-IAT variance, and criterion sensitivity.

Between-sample variance

Between-sample variance occurs when average response times and error rates vary between participants and thus between the 10,000 samples. When this happens, we cannot reliably infer stimulus validity because we do not know whether a new sample of 100 participants will be able to categorize the stimuli with similar speed and accuracy.

A possible cause of the between-sample variance found in this study is the unknown characteristics of our participant sample. Although we selected 18- to 25-year-old US participants, within that sample we did not further distinguish between potentially relevant characteristics. The unknown participant characteristics of our (sub-)populations may have caused the between-sample variance seen in the participant responses.

To illustrate why participant characteristics and/or sub-populations may have caused between-sample variance, consider the Asian-American-IAT. Within the sample of 18- to 25–year-old US participants, there were likely some of Asian-American descent. Van Ravenzwaaij et al. (2011) showed that the presence/absence of bias (i.e., D_{IAT}) depended on in- and outgroup membership. Bias was present when the opposing categories contained in- or outgroup members, but disappeared when participants categorized names of two opposing outgroups. A randomly selected sample with only participants of Asian-American descent (in-group) therefore likely performed differently than a mixed- or non-Asian-American sample (outgroup). Unknown participant characteristics, such as in- and outgroup membership, may therefore have caused between-sample variance, in turn causing unreliable validity estimates.

Between-IAT variance

The between-IAT variance seen in the contextual differences analyses complicates interpretation of unreliable estimates even further. Stimulus validity differs not only between samples within the same IAT but also between IAT contexts. Although the current analyses do not provide insights as to why contextual differences occurred, the most likely causes are participant and/or stimulus characteristics.

The Race-IAT consistently under performing in terms of RT validity may have been caused by participant characteristics. The Race-IAT is the most popular IAT, which is often included in introductory Psychology courses. If participants completed the Race-IAT during a lecture, this subset of participants may have different characteristics than participants that completed other IATs from home/the lab. Possible influential characteristics are being distracted while completing the IAT in class – causing longer response times. Or the increased pressure to provide socially desirable responses because of the presence of peers. Any of these participant characteristics, when different from other IATs could have caused between-IAT variance.

The Arab-IAT is another interesting example, because here it appears that stimuli characteristics rather than participant characteristics have caused between-IAT variance. Not the characteristics of the "Good"/"Bad" adjectives, which were used across the nine/ten analysed IATs, but the characteristics of the other two categories ("Arab Muslims"; "Other People"). Almost all stimuli in those categories were invalid, possibly due to the outgroup effect (van Ravenzwaaij et al., 2011) discussed earlier. It seems plausible that the US participants' unfamiliarity with "Arab"/"Other" names has affected the "Good"/"Bad" adjectives, which were answered incorrectly more often than those same stimuli were in other IATs.

Interpretation of unreliable validity estimates is thus complicated by the fact that unreliability differs between IATs. Furthermore, research is necessary to determine why and when stimulus validity differs between contexts. Without clear explanations, it appears unfounded to remove “Good”/“Bad” adjectives in only the one/two out of ten IATs in which the stimulus was invalid.

Criterion sensitivity

The third cause of between-sample variance is a difference in the ‘performance’ of the RT and ERROR validation criterion. Applying only the ERROR criterion would allow for 530 stimuli to remain included, whereas the RT criterion would only include 58 stimuli. In most cases, the RT criterion is more sensitive (i.e., concludes more invalidity) than the ERROR criterion. Yet, in some cases, the ERROR criterion did not conclude stimulus invalidity where the RT criterion did (e.g., “Grief,” “Humiliate,” and “Selfish” in the Race-IAT, Figure 5).

The difference in sensitivity is caused by a mismatch between the average response distributions and the absolute cut-offs. The ERROR criterion is less sensitive because the absolute cut-off (<10%) is higher than the average error rates of most participant samples (e.g., ±5% in the Age-IAT). The RT criterion (<800 ms) instead is more sensitive because the cut-off is placed around the average response time. The RT criterion therefore divides the centre of a normal response distribution into an unreliable split of ‘valid’ and ‘invalid’ estimates.

The sensitivity of the RT criterion is affected by the response time variances and the absolute cut-off. In the current analyses, the response time variances may be slightly larger because we retained reaction times which were faster than 400 ms (Raw data section). These ‘too fast’ responses hold valuable information. However, they also increase response time variances and decrease average response times. The actual sensitivity of the RT criterion may therefore be slightly overestimated in the current analyses.

The sensitivity of the RT criterion is also depended on the utilized cut-offs. Had we chosen the lower boundary suggested by Greenwald et al. (2021) – 600 ms instead of 800 ms – then the RT validity estimates would be less ‘unreliable’ and more consistently ‘invalid’. At the same time, higher RT boundaries would result in more reliably valid stimuli. For example, all stimuli in the Gender-Career-IAT would be deemed reliably valid if the RT criterion was set to <1200 ms.[10] However, such a high RT criterion is undesirable given that IAT responses are assumed to reflect automatic response tendencies (Greenwald et al., 1998). Automatic responses are typically fast; substantially less than 1000 ms on average (Botvinick et al., 2001; De Houwer, 2001; MacLeod, 1991; Ridderinkhof et al., 2021).

Greenwald et al. (2021) did not provide theoretical or empirical foundations for choosing the specific absolute cut-off boundaries. It is therefore unclear whether ‘performance’ differences are expected and perhaps even desired. If one assumes that participants prefer speed over accuracy, then the RT criterion should indeed be more sensitive to higher responses times. From that perspective, Greenwald et al. (2021)'s validation criteria are doing what they should be doing: penalizing all items with too slow responses ($N = 142$; 15.38%).

However, if one assumes that participants prefer to be accurate rather than fast, the RT criterion should be more lenient to slow responses because answering accurately requires additional time (for a detailed overview of the speed-accuracy trade-off, see Heitz, 2014). Longer response times are clearly necessary when stimuli are unfamiliar and/or contain cross-category associations. For example, the images of older presidents (e.g., Nixon) in the President-IAT were unreliable, whereas images of Trump were valid. The stimuli “Michelle” and “Daniel” stood out in the Gender-Career IAT because the average response times were considerably higher than other Male/Female names. This could reflect that common derivatives exist for the opposite genders (“Michel” & “Danielle”).

[10] The consequence of different RT boundaries can be globally inferred from the Figures (e.g., Figure 3). Detailed statistics can also be produced by re-running our analyses with different parameters (code: https://osf.io/dw23y/).

In sum, differences in criterion sensitivity caused between-sample variances because of the dichotomous classification of normally distributed data. Unreliability estimates can be decreased by changing the absolute cut-offs, but it is not clear how they should be changed. If one wants to validate items without making assumptions about the speed-accuracy trade-off, then relative-validation criteria may prove useful.

Relative and/or absolute validation criteria

Greenwald et al. (2021)'s validation criteria are absolute cut-offs in the sense that they are 'one size fits all'. The same validation criteria are applied to all stimuli, all stimulus types, all participants, and all IATs. The benefit of absolute validation criteria is the ability to compare because they are the only way to detect that an entire category/IAT is performing below expectation compared to 'peers'. Inferences about the quality of entire IATs, such as the Arab-IAT described above, would not be possible without criteria that apply to all IATs.

The issue with a 'one-size-fits-all' approach however is it does not account for non-problematic causes of variance. For example, the slower response times in the Race-IAT need not be problematic as long as the delay is constant across stimuli. D_{IAT}, the IAT's outcome measure, is ultimately a standardized difference score. Computing D_{IAT} for systematically longer response times therefore does not cause a problem. Longer responses do however cause a problem for stimulus validation. With a relative-validation criterion, one would first determine what is considered a normal response – in its simplest form akin to z-score outlier detection – and then invalidate stimuli from there. That way, the systematically longer response times are accounted for *before* determining stimulus (in)validity.

Relative-validation criteria would also resolve the differences in sensitivity between the RT and ERROR validation criteria. The RT criterion would become less sensitive, because it would account for the fact that the average response time is generally around the absolute cut-off of 800 ms. For 'most' people to meet the RT criterion, the cut-off should thus be somewhat higher. The ERROR criterion will simultaneously become slightly more tuned to stimuli that perform worse than expected, even though the error rates may not have exceeded the absolute cut-off of 10%.

Importantly, with relative-validation criteria, some stimuli that currently have gone undetected will then be classed as invalid. Examples include "Michelle" and "Daniel" in the Gender-Career IAT (see Figure 3), but also "Good"/"Bad" adjectives such as "Humiliate" where the average error rate was higher than 10% in 53.25% of the samples (Race-IAT, Figure 5). Relative-validation criteria would be able to signal individual outliers – but not consistent under performance. For example, comparisons of stimulus validity across stimulus types – which signalled issues with the Native-American-IAT – would not be possible.

Taken together, our research findings advocate for the use of both absolute- and relative-validation criteria, where absolute criteria could be sufficient if tuned to participant and/or stimulus characteristics. Despite our diligent attempts, we have yet to come across research that empirically tested absolute- or relative-validation criteria. Future research could therefore benefit from determining when response times should be considered 'implicit' or 'automatic'. In addition, future research should explore when and why between-sample variance occurs, so that we know when absolute- or relative-validation criteria are preferred.

Practical suggestions

So far, we have primarily focused on the theoretical implications of our research. But even though the validation criteria themselves are still up for discussion, our findings already allow for some practical recommendations.

We first recommend that researchers carefully read through Greenwald et al. (2021)'s "Best Practices." Some of the stimuli that stood out in our analyses would not have been selected if all best practices had been implemented. For example, "Michelle" and "Daniel" would likely not have been selected due to their idiosyncratic relationship with the opposing gender category (recommendation A7). Greenwald et al. (2021) also offer practical recommendations for administering IATs, which are not covered in the present article.

Our results indicate that some adjectives were consistently invalid. The second recommendation is thus to avoid use of stimuli, which were invalid across IATs. Based on RT validation, we recommend against using 'Terrific', 'Humiliate', 'Triumph', 'Selfish', and 'Bothersome'. These stimuli were invalid in all nine/ten IATs in which they occurred. Based on the ERROR validation, we recommend against using 'Terrific', which was again invalid in all nine/ten IATs in which it occurred. We also recommend to exclude stimuli, which were not invalid, but attained considerably lower validity than most other stimuli: 'Humiliate', 'Selfish', 'Triumph', and 'Grief'. For "Good"/"Bad" adjectives that are safe to include, please see the full-sized plots in the Online Appendix.

Our recommendation against some of the adjectives contradicts Axt et al. (2021) who concluded that all adjectives were safe to use. They compared the same adjectives, in similar Project Implicit data, but did not find between-IAT differences. Axt et al. (2021) however conducted leave-1-out analyses and explored the effect of individual stimuli on D_{IAT}. Where they analysed data from Blocks 3, 4, 6, and 7, we instead looked at raw response data from Blocks 1 and 2. Therefore, the findings appear contradictory, but because different data and outcome measures were used, it is technically possible that both conclusions are true. This would however imply a need for future research to determine which response data should be used to validate stimuli.

Our third recommendation is to create one IAT version, which is administered to the entire participant population. Axt et al. (2021) suggested that any of the evaluated adjectives may be randomly selected for inclusion. This is a common approach, where stimuli are randomly drawn from a larger pool, causing each participant to see a different version of the 'same' IAT. The between-sample variability suggests that this is not a sensible approach. Reliably estimating stimulus validity is difficult enough in samples of 100 participants, let alone if the number of participants per stimulus is reduced even further. The reduction of *N*, while keeping the variance stable, decreases power, which is a decreased chance of detecting true stimulus (in)validity.

Presenting participants with randomized versions of IATs is also a bad idea when we consider that the Arab-IAT attained higher error rates for the "Good"/"Bad" adjectives than the other nine/ten IATs. Closer inspection suggested that the invalidity of the two other categories decreased validity of the "Good"/"Bad" adjectives. The reverse could also be true, but would be difficult to detect if only some participants received problematic combinations of stimuli, whereas others did not.

Our fourth recommendation follows from the stimulus type analyses. If one aims for fast response times, then we can recommend using images as exemplars. On average, images see the lowest error rates and the fastest responses times – as is evident from the intercept estimates of the mixed-effects models. The results however also clearly show that some images are unsuitable for use in IATs. For example, the President-IAT shows signs of unfamiliarity for the images depicting older presidents (e.g., Nixon) but not for images depicting recent presidents (e.g., Trump). Ratliff & Smith (2021) also discussed a need for updating the Race-IAT stimuli, mainly due to poor image quality. They however questioned whether the small (validity) gains outweigh the costs of breaking a very extensive longitudinal data chain. In choosing any stimulus, it is thus good to also consider longevity.

Our final recommendation considers the debate of pilot- versus post-hoc testing. Greenwald et al. (2021) proposed that a small *pilot* sample would be sufficient to determine which stimuli should be in- or excluded. Our *post-hoc* analyses however show that two random samples of 100 can vary drastically in average response times and error rates. The between-sample variance is large enough to suggest that pilot results would not necessarily transfer to experimental populations. Post-hoc validation ensures that validation is performed for the participants for which D_{IAT}s are then computed. Post-hoc validation ensures that any (unknown) participant characteristics are accounted for, especially when relative-validation criteria are applied. We thus recommend post-hoc validation analyses, even if one only explores the data but does not (yet) apply absolute- or relative cut-offs. Although we advocate the need for post-hoc validation analyses, that is not to say that pilot-testing lacks value.

To summarize, we recommend applying Greenwald et al. (2021)'s best practices, excluding some adjectives, creating standardized IAT versions, and post-hoc validation analyses.

CONCLUSION

We explored the practical and theoretical utility of Greenwald et al. (2021)'s proposed validation criteria. They suggested that stimuli should only be included when they are easy to classify – translating to rapid (response times <800 ms) and accurate (error < 10%) participant responses (p. 7). Three sets of analyses show that the validation criteria are easily applied to the raw response data of 18–25–year-old US participants. Applying Greenwald et al. (2021)'s validation criteria signals that 94.15% of stimuli across 15 IATS should be removed. The presence of between-sample and between-IAT variance however warrants more nuanced interpretations. We therefore advocate the need for future research, which explores the causes of variance and the possibility of relative and/or stimulus-type-specific validation criteria.

AUTHOR CONTRIBUTIONS

Sally A. M. Hogenboom: Conceptualization; formal analysis; methodology; validation; visualization; writing – original draft; writing – review and editing. **Katrin Schulz:** Conceptualization; funding acquisition; supervision; writing – review and editing. **Leendert van Maanen:** Conceptualization; methodology; supervision; validation; visualization; writing – review and editing.

ACKNOWLEDGEMENTS

This publication is a part of the project “The biased reality of online media – Using stereotypes to make media manipulation visible” (with project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which is financed by the Dutch Research Council (NWO). We further want to acknowledge that this research would not be possible without the research efforts and open data policies of Project Implicit (https://www.projectimplicit.net/).

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest to report.

OPEN RESEARCH BADGES

This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at https://osf.io/y9hiq/ [data] and https://osf.io/dw23y [analysis, code, results].

DATA AVAILABILITY STATEMENT

This research utilizes data collected by Project Implicit (https://www.projectimplicit.net/). All data and the analyses scripts to support our findings are available via the Open Science Framework (https://osf.io/dw23y/).

ORCID

Sally A. M. Hogenboom https://orcid.org/0000-0003-3222-0019
Katrin Schulz https://orcid.org/0000-0001-8573-9768
Leendert van Maanen https://orcid.org/0000-0001-9120-1075

REFERENCES

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). *Rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bickel, P. J., Götze, F., & van Zwet, W. R. (2012). Resampling fewer than n observations: Gains, losses, and remedies for losses. In S. van de Geer & M. Wegkamp (Eds.), *Selected works of Willem van Zwet* (pp. 267–297). Springer. https://doi.org/10.1007/978-1-4614-1314-1_17

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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.I08.3.624

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the implicit association test. *Journal of Personal and Social Psychology*, *81*(5), 760–773. https://doi.org/10.1037/0022-3514.81.5.760

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Darling-Hammond, S., Michaels, E. K., Allen, A. M., Chae, D. H., Thomas, M. D., Nguyen, T. T., Mujahid, M. M., & Johnson, R. C. (2020). After "the China virus" went viral: Racially charged coronavirus coverage and trends in bias against Asian Americans. *Health Education & Behavior*, *47*(6), 870–879. https://doi.org/10.1177/1090198120957949

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

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303–315. https://doi.org/10.1086/209351

Garimella, A., Banea, C., & Mihalcea, R. (2017). Demographic-aware word associations. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2285–2295. https://doi.org/10.18653/v1/D17-1242

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

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology/Psychologie Canadienne*, *50*(3), 141–150. https://doi.org/10.1037/a0013848

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. H., Jost, J., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B., … Wiers, R. (2020). *The Implicit Association Test at age 20: What is known and what is not known about implicit bias*. PsyArXiv. https://doi.org/10.31234/osf.io/bf97c

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., … Wiers, R. W. (2021). Best research practices for using the implicit association test. *Behavior Research Methods*, *54*, 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

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. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

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

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. https://www.jstatsoft.org/v40/i03/

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General*, *148*(6), 1022–1040. https://doi.org/10.1037/xge0000623

Heiphetz, L., Spelke, E. S., & Banaji, M. R. (2013). Patterns of implicit and explicit attitudes in children and adults: Tests in the domain of religion. *Journal of Experimental Psychology. General*, *142*(3), 864–879. https://doi.org/10.1037/a0029714

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnins.2014.00150/full

Hester, J. (2020). *Glue: Interpreted string literals*. https://CRAN.R-project.org/package=glue

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

Hope, R. M. (2013). *Rmisc: Rmisc: Ryan miscellaneous*. https://CRAN.R-project.org/package=Rmisc

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychonomic Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Miller, S. (2018). Mixed Effects Modeling Tips: Use a Fast Optimizer, but Perform Optimizer Checks [Blog]. In *Steven V. Miller*. http://svmiller.com/blog/2018/06/mixed-effects-models-optimizer-checks/

Müller, K. (2020). *Here: A simpler way to find your files*. https://CRAN.R-project.org/package=here

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. https://doi.org/10.1177/0146167204271418

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

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. 39.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2022). *Foreign: Read data stored by 'minitab', 's', 'SAS', 'SPSS', 'stata', 'systat', 'weka', 'dBase'*, … https://CRAN.R-project.org/package=foreign

Ratliff, K. A., & Smith, C. T. (2021). Lessons from two decades of project implicit. In *The Cambridge handbook of implicit bias and racism*. Cambridge University Press.

Ravary, A., Baldwin, M. W., & Bartz, J. A. (2019). Shaping the body politic: Mass media fat-shaming affects implicit anti-fat attitudes. *Personality and Social Psychology Bulletin*, *45*(11), 1580–1589. https://doi.org/10.1177/0146167219838550

Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P. M., Bashore, T. R., & van der Molen, M. W. (2021). The arrow of time: Advancing insights into action control from the arrow version of the Eriksen flanker task. *Attention, Perception, & Psychophysics*, *83*(2), 700–721. https://doi.org/10.3758/s13414-020-02167-z

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the implicit association test: The recoding-free implicit association test (IAT-RF). *Quarterly Journal of Experimental Psychology*, *62*(1), 84–98. https://doi.org/10.1080/17470210701822975

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Steffens, M. (2005). Implicit and explicit attitudes towards lesbians and gay men. *Journal of Homosexuality*, *49*(2), 39–66. https://doi.org/10.1300/J082v49n02_03

Steffens, M., & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the implicit association test. *Zeitschrift Für Experimentelle Psychologie: Organ Der Deutschen Gesellschaft Für Psychologie*, *48*, 123–134. https://doi.org/10.1026//0949-3946.48.2.123

Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: Direct replication of the predictive validity of the suicide. *Psychological Science*, *31*(1), 65–74. https://doi.org/10.1177/0956797619893062

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the name-race implicit association test measure racial prejudice? *Experimental Psychology*, *58*(4), 271–277. https://doi.org/10.1027/1618-3169/a000093

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. https://CRAN.R-project.org/package=readr

Wickham, H., & Seidel, D. (2020). *Scales: Scale functions for visualization*. https://CRAN.R-project.org/package=scales

Wilke, C. O. (2020a). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot

Wilke, C. O. (2020b). *Ggtext: Improved text rendering support for 'ggplot2'*. https://CRAN.R-project.org/package=ggtext

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Xie, Y. (2016). *Bookdown: Authoring books and technical documents with R markdown*. Hall/CRC. https://github.com/rstudio/bookdown

Xie, Y. (2022). *Knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/

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

APPENDIX A

Implicit measures of attitudes, among which the IAT, aim to surpass the social desirability bias often associated with self-report measures (Fazio et al., 1995; Greenwald et al., 1998). The social desirability bias is the participants' tendency to provide responses that are socially acceptable rather than a reflection of their true attitudes (Maccoby & Maccoby, 1954 in Fisher, 1993). Implicit measures surpass this bias[11] by measuring attitudes indirectly, thereby revealing biases that were less pronounced or non-existent when measured with explicit/direct self-report questionnaires (Fazio et al., 1995; Greenwald et al., 1998; Hofmann et al., 2005). The IAT, for example, measures bias indirectly by comparing response times across a 7-block categorization task (Greenwald et al., 1998, 2003). Because implicit measures surpass social desirability biases they are used most often in contexts where such biases are likely to occur. Examples include measuring sexual attraction to children (Babchishin et al., 2013), racial biases towards Asian-Americans after terming COVID-19 the "China Virus" (Darling-Hammond et al., 2020), or measuring an individuals' suicidal thoughts (Tello et al., 2020).

The IAT measures the association strength (i.e., bias) between target-categories and attitude- or stereotype-categories. The target-categories (e.g., "Christian" vs. "Muslim," Heiphetz et al., 2013) are paired with attitude-categories (e.g., "Pleasant" vs. "Unpleasant," Greenwald et al., 1998) to reveal differences in association strength between two sets of categories (e.g., Heterosexual-Positive & Gay-Negative, Steffens, 2005). Figure A1 illustrates the procedure of the *Gender-Career IAT* (GC-IAT), which is aimed at understanding implicit attitudes towards traditional gender roles by measuring association strengths between the categories Career/Family (target-categories) and Male/Female (stereotype-categories). Each of the categories is represented by multiple exemplars (together called the stimuli), which provide information about the overarching category. For example, the category "Female" is represented by the exemplars "Rebecca," "Michelle," "Julia," "Emily," and "Anna." Critically, the association strength between the categories is inferred from the participants' responses to the exemplars, and not directly from the responses to the categories themselves.

The categories, pictured on two sides of the screen, correspond with designated response keys (e.g., "E" = left; "I" = right). Participants sort the exemplars into the correct categories by pressing the corresponding response key. For each trial, response time (RT) and accuracy (incorrect/correct) are recorded. Note that the IATs evaluated in this research include a built-in-error penalty; participants must change an incorrect answer before the response time is recorded. In Blocks 1, 2, and 5, participants assign the exemplars into two opposing categories: the target-categories (Career/Family) or the stereotype-categories (Male/Female). Blocks 3/4 and 6/7 are the so-called critical blocks: participants sort exemplars into four categories that are paired into two response options (left and right). The pairing of the categories

[11] Among others, Hofmann et al. (2005) and Gawronski (2009) provide mixed evidence accounts of implicit measures effectively reducing social desirability biases.

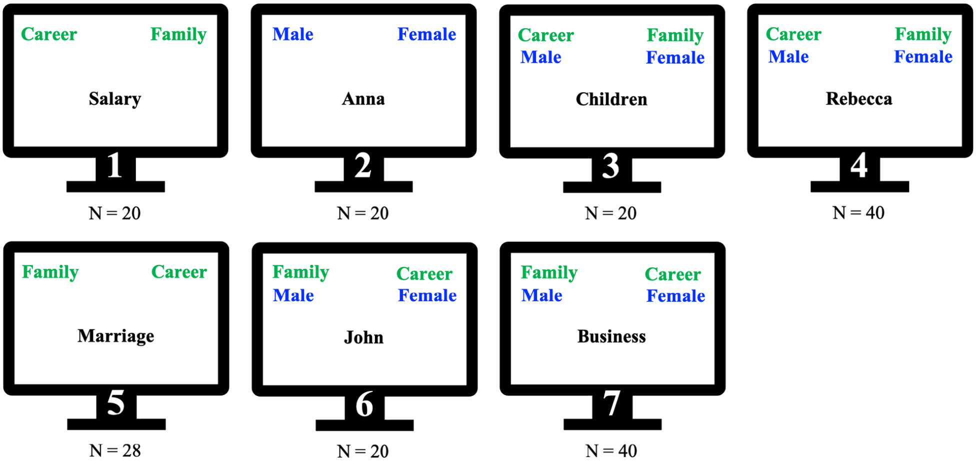


FIGURE A1 Schematic overview of the Gender-Career Implicit Association Test (GC-IAT). The IAT consists of 7 blocks where participants sort the exemplar (black) into categories (green or blue) by pressing the correct response key. This visualization does not include instruction screens and response-key instructions (e.g., left = "E" & right = "I"). The number of trials (*N*) differ across blocks and IATs due to variations in the number of exemplars per category. Adapted from the GC-IAT on Project Implicit (https://implicit.harvard.edu/implicit/Study?tid=-1).

on the left or right side of the screen determines whether a block is considered congruent or incongruent (i.e., compatible vs. incompatible). In the GC-IAT, the association between Career-Male (left) and Family-Female (right) is considered congruent because these pairings reflect the traditional gender roles, whereas pairings of Career-Female (left) and Family-Male (right) are considered incongruent (see Figure A1). The (in)congruent pairing of categories in the blocks 3/4 versus 6/7 are counter-balanced across participants to reduce the chances of order effects.

For each participant, RTs are used to compute the participants bias score (D_{IAT}). The participants' bias score expresses the association strength between the four categories (Nosek et al., 2005). In the case of the GC-IAT, a positive D_{IAT} indicates that the participant responded faster when the categories were paired Career-Male and Family-Female than when the categories were paired Career-Female and Family-Male. Note that the IAT is a parallel association task; all four categories are presented at once. We thus cannot infer a difference in association strength between two categories (e.g., Career-Male vs. Career-Female), but only the difference between the four categories in different pairings (Brendl et al., 2001). Central to our research is the fact that both the categories and exemplars exert an effect on D_{IAT} (Gast & Rothermund, 2010) and (in)appropriate stimuli selection directly effects the (direction of) the measured bias score.