Implicit Association Tests: Stimuli Validation from Participant Responses

Abstract

The Implicit Association Test (IAT, [Greenwald et al., 1998](#ref-GREENWALD1998)) is a popular instrument for measuring attitudes and (stereotypical) biases. The outcome measure is directly affected by a stimuli (categories and exemplars) and participant interaction. This interaction may cause undesirable stimulus effects – stimulus unfamiliarity and/or cross-category associations – which directly affect the direction and size of . Changes to the stimuli or the target population therefore warrant (re-)validation of the included stimuli to ensure that undesirable stimulus effects were not accidently introduced. Greenwald et al. ([2021](#ref-GREENWALD2021)) propose a concrete method for validating stimuli: appropriate stimuli should be familiar and easy to classify – translating to rapid (response times < 800 ms) and accurate (error < 10%) participant responses. We will conduct three analyses to explore the theoretical and practical utility of these proposed validation criteria. We first apply the proposed validation criteria to the data of 15 IATs that were available via Project Implicit. We then explore the context dependency of validity by comparing the validity of stimuli that are used across multiple IATs. Finally, we study the sensitivity of stimulus validity for different stimulus types.

*Keywords:* Implicit Association Test, Internal Validity

*Word count:* 6,825

Implicit Association Tests: Stimuli Validation from Participant Responses

# 1 Introduction

The *Implicit Association Test* (IAT, [Greenwald et al., 1998](#ref-GREENWALD1998)) 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 three years, the number of IAT studies increased from 3,608 (March 2019, [Greenwald et al., 2020](#ref-GREENWALD2020)) to 4,172 (February 2022): an average of 16 publications per month [[1]](#footnote-20).

Some of these publications raise concerns about the construct validity of the IAT (cf., [De Houwer, 2001](#ref-DEHOUWER2001); [Gawronski, 2009](#ref-GAWRONSKI2009)). For example, studies show low predictive validity ([Greenwald et al., 2009](#ref-GREENWALD2009)), low test-retest reliability ([Hehman et al., 2019](#ref-HEHMAN2019)), and a lack of discriminant validity ([Schimmack, 2021](#ref-SCHIMMACK2021)). A possible explanation for the IAT’s measurement issues is the lack of convergence in the utilized stimuli. Axt et al. ([2021](#ref-AXT2021)) 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](#ref-AXT2021)), are not only theoretical. Multiple studies show that stimulus choices directly affects the size and direction of the measured IAT bias (e.g., [Bluemke & Friese, 2006](#ref-BLUEMKE2006); [Steffens & Plewe, 2001](#ref-STEFFENS2001)).

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](#ref-GREENWALD2021)) 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](#ref-GREENWALD2021)). 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 will thus explore the theoretical and practical utility of these criteria as validation measures for IAT stimuli.

## 1.1 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 discuss specifically the IAT’s stimuli. A more detailed description of the entire IAT paradigm is available in Appendix 3.2.

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](#ref-HEIPHETZ2013)) or attitudes (e.g., “Pleasant” vs. “Unpleasant,” [Greenwald et al., 1998](#ref-GREENWALD1998)). An IAT contains two sets of opposing categories which together 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 ([Gast & Rothermund, 2010](#ref-GAST2010)) and (in)appropriate stimuli selection, as we will discuss in the section below, directly affects the (direction of) the measured bias score.

## 1.2 Undesirable Stimulus Effects

Although it is evident that the IAT is only as good as the included stimuli – Greenwald et al. (preprint: [2020](#ref-GREENWALD2020); article: [2021](#ref-GREENWALD2021)) are the only researchers to explicitly describe how stimuli should be selected to prevent undesirable stimulus effects. We focus specifically on undesirable stimulus effects which occur due to an interaction between stimuli characteristics and participant characteristics: stimulus unfamiliarity and cross-category associations.

### 1.2.1 The Issue of Stimulus Unfamiliarity.

Greenwald et al. ([2021](#ref-GREENWALD2021)) recommend that stimuli should be familiar to the participants ([2021](#ref-GREENWALD2021), Table 1: A1 - A2). When participants are unfamiliar with categories the IAT cannot be expected to measure bias, may cause spurious correlations, yield negative biases or null-effects ([Greenwald et al., 2020](#ref-GREENWALD2020); [Greenwald et al., 2021](#ref-GREENWALD2021)). Brendl et al. ([2001](#ref-BRENDL2001)) even showed that unfamiliar nonword 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](#ref-GREENWALD2021)).

(Un)familiarity with categories is however 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 of “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.

### 1.2.2 The Issue of Cross-Category Associations.

A second important recommendation by Greenwald et al. is to select stimuli that avoid cross-category associations ([2021](#ref-GREENWALD2021), 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: negation (“trust” & “distrust”), image patterns (all women smiling thus positive, all men frowning thus negative), and causation (“cancer” & “smoke” are both negative and have a cause-effect relationship). Each of these cross-category associations exist because the stimuli themselves have an additional property which allows them to be associated with multiple categories.

Two studies exemplify the effect of cross-category associations on the direction and size of [for more detailed overviews see; Greenwald et al. ([2021](#ref-GREENWALD2021)); Axt et al. ([2021](#ref-AXT2021))]. Steffens and Plewe ([2001](#ref-STEFFENS2001)) 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 where 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](#ref-BLUEMKE2006)). They did similar experiments where they manipulated the relationship between the target-categories (East- & West-German nouns and names), the attitude-categories (positive & negative nouns), and the participants. For example, the exemplar “Stasi” was used as a negative exemplar with a cross-category association to former East-Germany. Their experiments showed that IAT effects could be increased if the manipulations favored the in-group participants (West-Germans), whereas the IAT effects were decreased when the manipulations favored the out-group participants (East-Germans). These results suggest that the size of 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 which are unfamiliar or contain cross-category associations have the potential to change the direction and size of IAT effects (). 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’ and 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.

## 1.3 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](#ref-GREENWALD2021)) proposed two absolute cut-off criteria which 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 which do not meet these criteria should be “[…] discarded without further consideration.” (p.7, [Greenwald et al., 2021](#ref-GREENWALD2021)). Greenwald et al. ([2021](#ref-GREENWALD2021)) further suggest that these validation criteria should be applied to data from *pilot* subjects originating from the intended participant population. Validating the stimuli according to Greenwald et al. ([2021](#ref-GREENWALD2021))’s recommendations allow 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, over 2,000 publications[[2]](#footnote-26) utilized data collected by Project Implicit[[3]](#footnote-27): a website where anyone can take part in IAT research. The demographic characteristics of Project Implicit participants is unknown a-priori and is difficult to predict due to the substantial number of participants each year (e.g., > 15,000 participants for the Race-IAT in 2020; see section 2.1.2). 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 apply Greenwald et al. ([2021](#ref-GREENWALD2021))’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 apply Greenwald et al. ([2021](#ref-GREENWALD2021))’s proposed validation criteria as post-hoc analyses to Project Implicit data. In total we conduct three sets of analyse which together evaluate the theoretical and practical utility of the criteria as pilot- and post-hoc validation analyses.

## 1.4 Research Aims

The purpose of our research is to evaluate the proposed validation criteria that exemplars should elicit fast (RT < 800 ms) and accurate (< 10% error) participant responses ([Greenwald et al., 2021](#ref-GREENWALD2021)). We do so by applying these criteria across a large body of existing IAT data[[4]](#footnote-31) in three sets of analyses.

In the first analysis, we aim to gain an overview of how the proposed validation criteria perform across individual IATs. We do this by applying the validation criteria to the data of 15 IATs separately. For each IAT we create 10,000 independent samples of 100 participants and determine stimulus validity within each sample. The fluctuations in the validity judgments across 10,000 re-samples provide evidence of the reliability with which one can infer validity from the response data of a random sample of 100 participants.

In the second analysis, we focus 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 section 1.2). 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 provide a unique opportunity to explore the context dependency of stimulus validity because some stimuli are used across multiple IATs. For example, the Age-IAT and the Skin-Tone-IAT both use 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 explore the potential dependency of stimulus validity on IAT context.

In the third and final analysis we aim to determine the effect of stimulus type on stimulus validity. A closer look at the stimuli used within and across IATs show substantial differences with stimuli varying from verbal to visual representations. To illustrate, the Gender-Career IAT exclusively uses proper 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](#ref-GREENWALD2021)) are equally sensitive for different stimulus types. In other words, what is the effect of stimulus type on stimulus validity? This third analysis thereby explores whether the proposed validation criteria would cause specific stimulus types to become (in)validated more often than others.

## 1.5 Potential Implications

Each of the three analyses described above has important implications for existing and subsequent IAT research. Prior to conducting these analyses, akin to hypotheses formulation, we can think of general outcome scenarios and their potential implications. We will revisit each of these scenarios in the *Discussion*.

In the first analysis we explore the validity of stimuli within 15 individual IATs. In an optimal scenario all stimuli would be deemed valid. However, based on the pilot analyses reported in section 3.1 this appears implausible. A more likely scenario would be that at least some IATs will contain stimuli that are categorized as invalid. The implications of these 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](#ref-NOSEK2005)) 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 after the invalid data has 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 explore the context dependency of stimulus validation. Among the possible results 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 compare how the validation criteria apply across different stimulus types. 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 section 1.2). If we find invalid stimuli in our participant sample, this need not imply that all Project Implicit data is 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.

# 2 Methods

We prepared our manuscript with RMarkdown in R [Version 4.2.1; R Core Team ([2020](#ref-R-base))][[5]](#footnote-36) which had several benefits. We reduced research degrees of freedom ([Wicherts et al., 2016](#ref-WICHERTS2016)) by formalizing data preprocessing and exclusion (section 2.2), the analyses (section 2.3), and to some extent the results (section 3.1). We conducted pilot analyses (section 2.1) which we used to optimize all code, but we expect that minor code changes may be necessary due to data irregularities (e.g., variable names). Because the analysis pipeline was already scripted for the pilot analyses, git version control will allow anyone to track and verify (the need for) those changes. Finally, preparing the analyses in R and making the code publicly available fosters 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 (<https://osf.io/dw23y/?view_only=25b62f307a1349e7883549b473091483>) which is connected to a Github repository (*removed for review purposes*).

In section 2.1 we will first discuss the origin and nature of the data we base our analyses on. We will also describe the differences between the pilot data and the proposed full data set. In section 2.2 we will describe the exclusion criteria and how each of them affects the data with which we are working. We then describe in section 2.3 the intended plan of analyses: how we computed the (reliability) of the validation criteria and intend to compare stimulus validity across IATs and stimulus types (nouns, names, images). The full research plan was approved by the Ethics Review Board at the *name removed for review purposes* [[6]](#footnote-38).

## 2.1 Data

For our analyses we will use data provided by Project Implicit[[7]](#footnote-39). Project Implicit is a large-scale data collection project which has collected IAT web responses since 2002 [Greenwald et al. ([2003](#ref-GREENWALD2003))][[8]](#footnote-40). Visitors to the website must agree to the terms and conditions before they have access to IATs about presidential preferences, body-weight attitudes, race and more.

The data that Project Implicit provides comes in two forms: compressed and raw. The compressed data contains one row of information per participant and includes information on demographics (e.g., age, occupation), IAT results (e.g., ), and explicit associations (i.e., self-report questions). It is freely available via Project Implicits’ Open Science Framework (OSF; <https://osf.io/y9hiq/>) for 16 IATs from 2002 until 2020[[9]](#footnote-42). The compressed data have primarily been used by researchers to determine group-level biases. For example, Charlesworth and Banaji ([2019](#ref-CHARLESWORTH2019)) performed trend analyses of biases from 2007 - 2016, Darling-Hammond et al. ([2020](#ref-DARLING-HAMMOND2020)) explored the effects of the Corona virus on Asian biases from 2007 - 2020, and Ravary et al. ([2019](#ref-RAVARY2019)) found that “fat-shaming” incidents predicted spikes in the biases detected with the body-weight IAT.

The raw data is available only upon request. It contains 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](#ref-ROTHERMUND2009)), determine the minimal number of exemplars ([Nosek et al., 2005](#ref-NOSEK2005)), and determine the effects of random stimulus variation ([Wolsiefer et al., 2017](#ref-WOLSIEFER2017)). The raw data and the compressed data can be linked via session\_id; a unique identifier for each started IAT session.

### 2.1.1 Pilot Data.

For the pilot analyses we worked with data from the *Gender-Career IAT* (GC-IAT). The compressed data of the GC-IAT was freely available via Project Implicit’ OSF page (<https://osf.io/abxq7/>). We received the raw response data from November and December 2019 from Project Implicit on the 3rd of 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 section 2.2.2), the compressed pilot data consisted of 8,509 rows of data. 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 a single person contributed a significant number of sessions. We further discuss the issue of repeated measurement in the exclusion criteria (section 2.2).

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 section 2.2.1), the raw pilot data consisted of 1,598,540 rows of data (responses/trials).

### 2.1.2 Full Data.

The full data we will utilize for this research comes from November 2020; the last full month for which data was available for all conventional IATs [[10]](#footnote-45). We will analyze the data from 15 IATs: the Age-IAT, Arab-IAT, Asian-IAT, Disability-IAT, Gender-Career-IAT, Gender-Science-IAT, Native-American-IAT, President-IAT, Race-IAT, Religion-IAT, Sexuality-IAT, Skin-IAT, Transgender-IAT, Weapons-IAT, and Weight-IAT.

The *compressed* data for the year 2020 is already publicly available for all IATs allowing us to apply the preprocessing steps mentioned above and estimate the number of eligible participants after applying the exclusion criteria (see section 2.2). The Race-IAT has the most eligible participants (N = 18,178) and the Asian-IAT the least (N = 491; = 5,336; = 4,377). The *raw* data has not yet been attained and will only be shared by Project Implicit after Stage 1 acceptance. Consequently, the final number of participants included in our analyses may be slightly lower than reported as participants may still be excluded based on raw response patterns.

## 2.2 Exclusion Criteria

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

### 2.2.1 Raw Data.

The raw data includes trial-by-trial information such as response times and errors. Previous research excluded slow and/or recoded fast response times. In line with guidelines set forth by Greenwald et al. ([2021](#ref-GREENWALD2021); [2003](#ref-GREENWALD2003)) we will apply an upper limit excluding trials where the response time is larger than 10,000 ms. We will 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 will exclude responses with a response time of zero milliseconds as these are likely the cause of technical malfunction. Note that the trials are removed, but the participants themselves are not excluded.

Contrary to suggested guidelines, in the present research we will exclude participants based on error percentages. Greenwald et al. ([2003](#ref-GREENWALD2003)) 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.

### 2.2.2 Compressed Data.

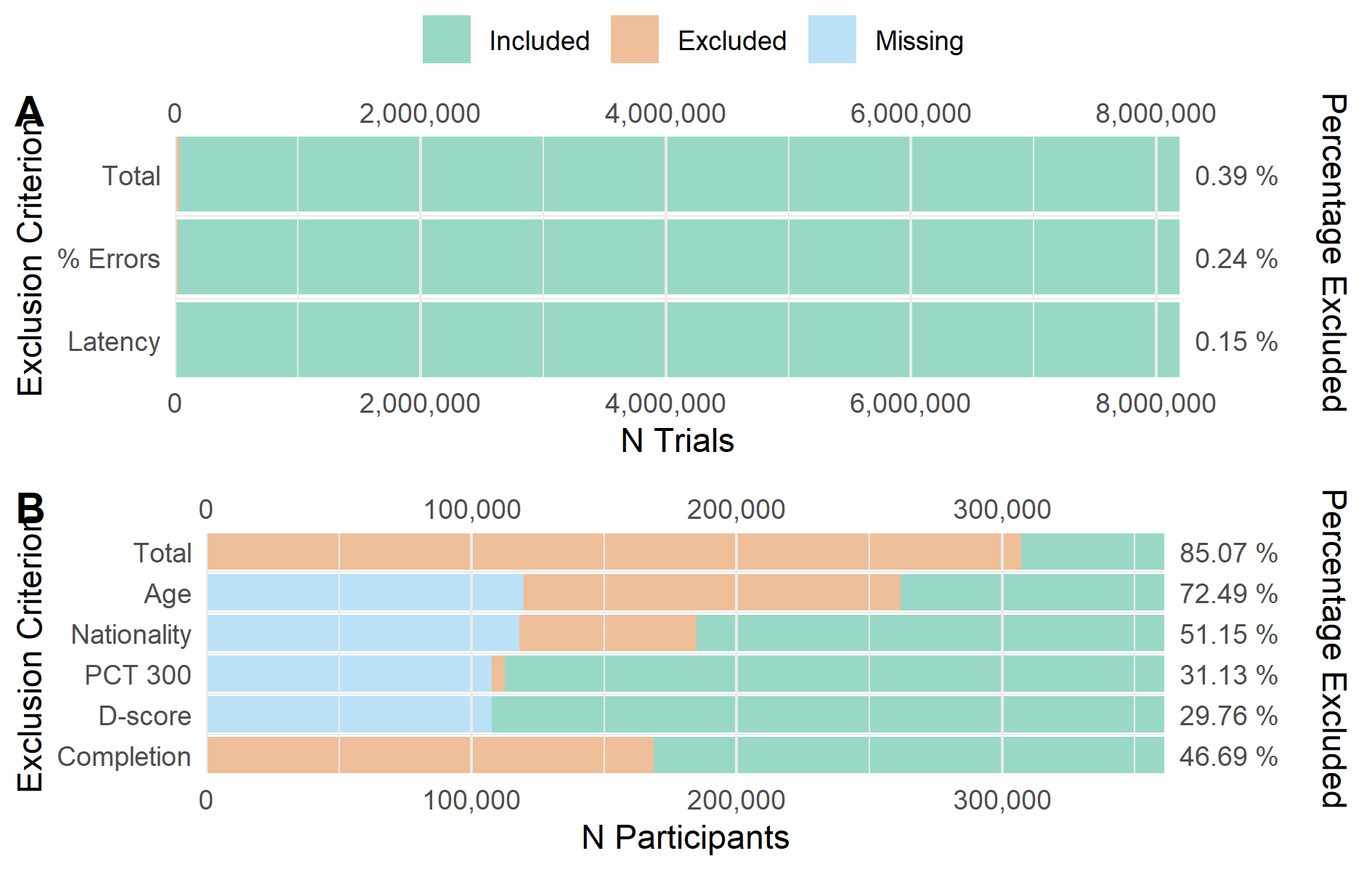
The compressed data contains information on summary statistics of the IAT, demographics, and answers to explicit questions (removed during data preprocessing). Participants who did not complete the full IAT will be removed from further analyses based on missing and/or an incomplete session\_status.

Participants were at times excluded who self-reported to have prior IAT experience. In the new guidelines, Greenwald et al. ([2021](#ref-GREENWALD2021)) state that the IAT retains its’ measurement properties with repeated measurement, although evidence indicates more polarized results at the first IAT measurement (section 2–4). We thus opted to refrain from exclusion based on prior IAT experience.

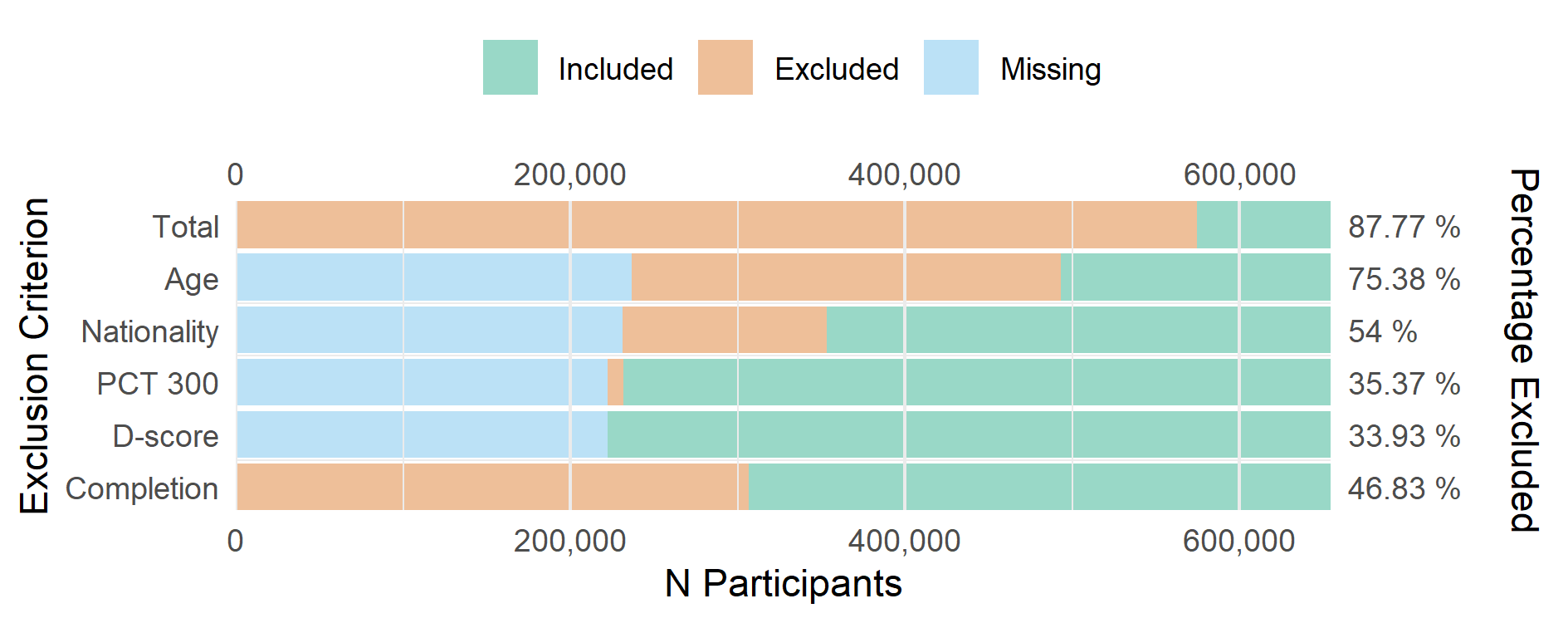
Greenwald et al. ([2003](#ref-GREENWALD2003)) showed the statistical benefits of removing participants who made too many fast responses. Participants who responded faster than 300 ms on more than 10% of the trials will be excluded from subsequent analyses (Step 5 in Appendix B, [Greenwald et al., 2021](#ref-GREENWALD2021)).

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 regards 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 will thus exclude participants that did not reside in, nor have citizenship of, the United States of America (USA). We will further exclude 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 – 18; <https://www.ageofconsent.net/states>). We will exclude participants who self-reported to be younger than 18 years old at the time of IAT completion. Greenwald et al. ([2021](#ref-GREENWALD2021)) further proposed that the validity of exemplars may be derived from examination of the data of “young adult subjects” (p. 7). We define “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 data. (A) The effects of the exclusion criteria on the ***raw*** pilot 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., ) 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 ***compressed*** 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 *raw* data were excluded from the *compressed* data, and vice versa. After mutual exclusion, 8,509 participants were eligible for analyses.



*Figure* *2.*  The aggregated effects of the exclusion criteria when applied to the ***compressed*** data of 15 IATs. Even though exclusions based on the raw data (e.g., percentage of errors) are not included, it is noteworthy that the exclusion percentages are comparable to the pilot analyses of the GC-IAT (2019; Figure 1). *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., ) 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.

## 2.3 Analysis Plan

In this study we will implement the validation criteria suggested by Greenwald et al. ([2021](#ref-GREENWALD2021)), which we discuss in more detail in section 2.3.1. We have already formalized these first analyses (section 2.3.2) and the results (section 3.1) in R code based on the pilot data from the GC-IAT. These analyses are therefore ready to be applied to the data of the 15 IATs included in the full data set. The validity estimates per stimulus that follow from the first set of analyses are used for two additional analyses. In section 2.3.3 we discuss how we will explore the dependency of stimulus validity on the IAT context in which a stimulus occurs. We then discuss in section 2.3.4 how we will establish the effect of stimulus type (words; images) on stimulus validity.

### 2.3.1 Validation Criteria.

We first focus on establishing the validity of stimuli (i.e., exemplars) within an IAT. Greenwald et al. ([2021](#ref-GREENWALD2021)) proposed:

A judgment 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](#ref-GREENWALD2021), p. 7)

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.

### 2.3.2 Sample Size and Bootstrapping.

Greenwald et al. ([2021](#ref-GREENWALD2021)) 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 inferred from the research population of interest rather than an unrelated pilot sample or prior research populations. This is especially relevant when we consider that (un)familiarity of stimuli and cross-category associations - as discussed in section 1.2 - are likely to differ among populations. Garimella et al. ([2017](#ref-GARIMELLA2017)), for example, showed that free-associations differed depending on gender (Male/Female) and location (USA/India). The cue “bath” was, for example, 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](#ref-GREENWALD2009)) included 184 independent samples with an average of 81 participants ( = 141.53, = 9, = 1,386). Babchishin et al. ([2013](#ref-BABCHISHIN2013)) included 12 studies with an average of 66 participants ( = 24.39, = 38, = 113). Finally, Oswald et al. ([2013](#ref-OSWALD2013)) included 97 independent samples with an average of 65 participants ( = 113.66, = 12, = 1,057). To err on the side of caution, the following analyses rely on a sample size that is somewhat higher than the averages of the reported meta-analyses: 100 participants. In doing so we increase the chances of finding true effects within a sample (i.e., power) while staying within reach of what is feasible for (future) experimental studies.

The data provided by Project Implicit, however, is much more extensive than a sample of 100 participants (e.g., = 8,509). The size of this data provides 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 will conduct 10,000 *m* out of *n* bootstraps ([Bickel et al., 2012](#ref-BICKEL2012)) 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*) will differ across IATs from 491 to 18,178.

Within each of the 10,000 samples we will determine the average response time and error rate of the 100 participants per stimulus; classifying stimuli as valid if the average response time is less than 800 milliseconds and the error rate is less than 10%. In total, we will thus have 10,000 classifications of validity per stimulus which together provide a reliability estimate of stimulus validity. We define stimuli as reliably *valid* if the stimulus was judged valid in 95% or more of the samples. Similarly, a stimulus was classed as reliably *invalid* if the stimulus was judged 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 judgment.

We will determine whether stimuli are reliably valid for the response time and error rate criteria separately. This provides a better insight into which of the two criteria has the biggest impact on judgments of stimulus validity. However, Greenwald et al. ([2021](#ref-GREENWALD2021)) clearly describe that valid stimuli are rapidly (response time) *and* accurately (error rate) categorized. We thus also computed a final validity judgment 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 judgment of *unreliable* as we could not be sure about the validity of both criteria. These final validity judgments are used for the two following analyses where we compare validity across IATs.

### 2.3.3 Context Dependent Stimulus Validity.

The analyses described above will result in validity estimations for each stimulus in each IAT. These estimations can subsequently be used to compare validity of the same stimulus across different IATs. In this analysis we choose to examine only verbal stimuli that are identical. We may encounter derivatives of the same stem (e.g., “Joy” and “Joyous”), but will not include those in the current analyses. Such derivatives could introduce variations in stimulus validity across IATs due to unknown lexical-syntactic properties which would prevent clear inferences.

Without seeing the raw data, which we will only attain after Stage 1 acceptance, we cannot be sure of the number of repeated stimuli nor the number of times a stimulus is repeated across IATs. We however expect large differences in the number of IATs in which each of the stimuli is included because often only the attribute- and stereotype-categories are reused. We thus believe it will be unlikely that drawing inferences from statistical tests would be appropriate. Instead, we will adapt the descriptive results figure for the by-IAT validity (i.e., Figure 3 below) to a by-stimulus figure; providing clear visual comparisons tailored to each stimulus separately.

### 2.3.4 Stimulus Type and Validity.

The purpose of this analysis is to determine whether stimulus type influences stimulus validity. Ultimately, we are interested in predicting validity (*percentage\_valid*; continuous) from stimulus types (*stimulus\_type*; categorical). To study this, we plan a mixed-effects regression model analysis, in which we will include *percentage\_valid* as the dependent variable and *stimulus\_type* as a fixed-effect. However, as stimulus types are nested within IATs we also need to account for both between-IAT and within-IAT effects. We account for between-IAT effects by including *IAT* as a random-effect: each *IAT* will be fitted with a unique *percentage\_valid* intercept. To account for the possibility that stimulus types nested within IATs exert an effect on validity we include *stimulus\_type* as a random slope: the relationship between *percentage\_valid* and *stimulus\_type* can differ between IATs.

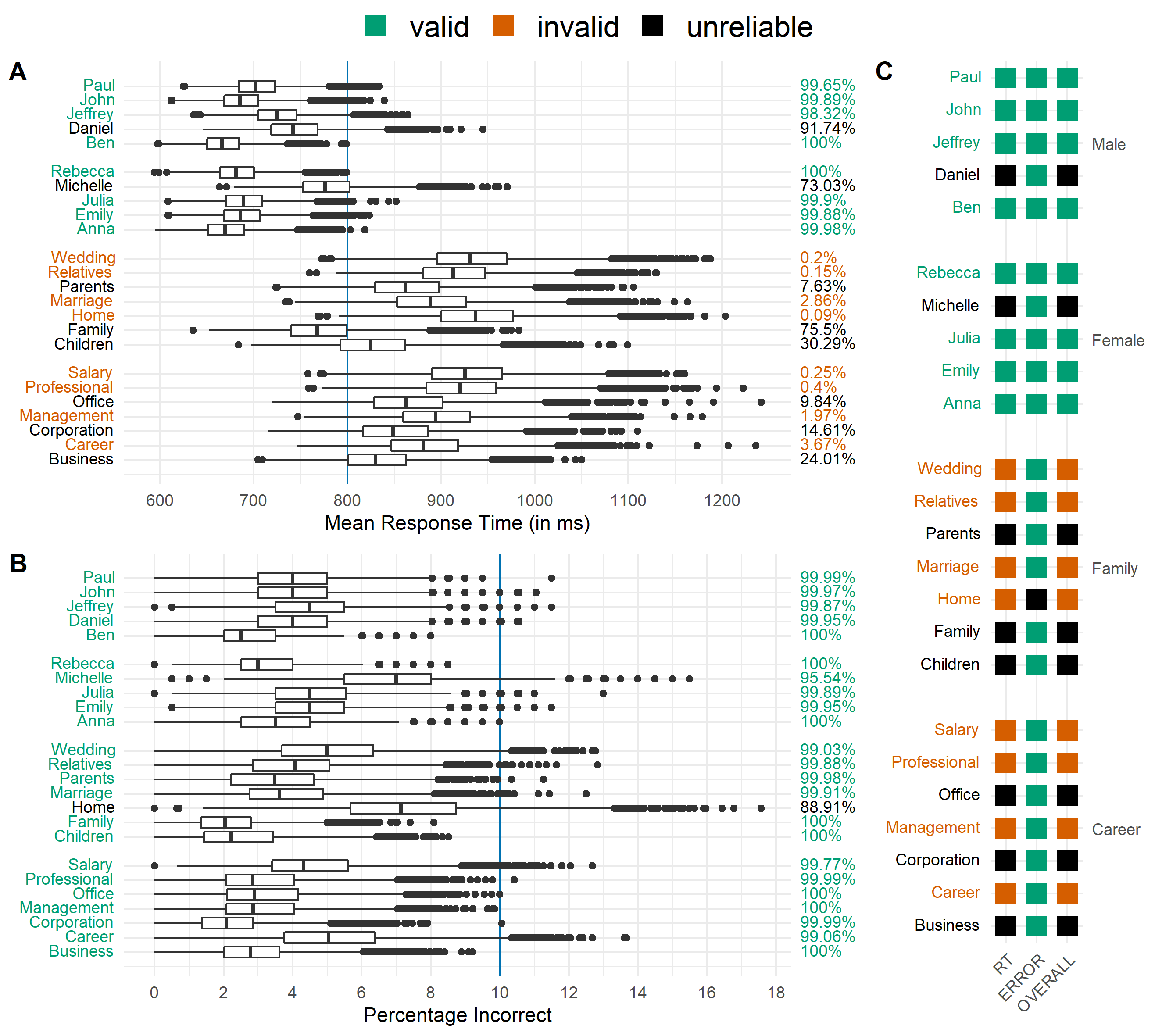
We will compute the following mixed-effects model using the lme4 ([Bates et al., 2015](#ref-R-lme4)) and lmerTest ([Kuznetsova et al., 2017](#ref-R-lmerTest)) packages with a BOBYQA optimizer ([Powell, 2009](#ref-POWELL2009)) and a maximum of 200,000 iterations ([Miller, 2018](#ref-MILLER2018)):

# 3 Results

## 3.1 Pilot

We conducted pilot analyses with the data of the GC-IAT to formalize the analyses and reports of the results for individual IATs. The results included here will be generated for each of the 15 IATs included in the final analysis. The pilot data included 8,509 participants who provided 1,598,540 responses overall and 340,100 responses in Block 1 and 2. The included participants were on average 20.67 years old ( = 2.43; 95% CI = 20.72, 20.62).

Greenwald et al. ([2021](#ref-GREENWALD2021)) propose that valid stimuli are those that are rapidly (RT < 800 ms) and accurately (errors < 10%) categorized (p. 7). 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 8 of the stimuli used in the GC-IAT were reliably valid (33.33%), 8 reliably invalid (33.33%), and 8 could not be reliably estimated (33.33%). With respect to the percentage of errors 23 of the stimuli used in the GC-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, 8 of the stimuli used in the GC-IAT were reliably valid (33.33%), 8 reliably invalid (33.33%), and 8 could not be reliably estimated (33.33%).



*Figure* *3.*  The validity estimates per stimulus of the Gender-Career IAT (November - December, 2019). **(A)** 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 judgment (right y-axis) and 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). **(B)** 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 percent of the trials were answered incorrectly (vertical blue line). **(C)** An overview of the validity judgments per exemplar based on average response times (RT) and percentage of errors (ERROR). A stimulus was classed as valid (OVERALL) if both criteria were also reliably valid, but as invalid if either criterion was reliably invalid. If the criteria were both unreliable, a stimulus was also classed as unreliable.

# 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*. <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>

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 Personalily 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 unobstrusive 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. <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*, 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. 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*. <https://doi.org/10.3758/s13428-021-01624-3>

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., 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>

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>

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>

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>

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. (2019). Forecasting a Fatal Decision: Direct Replication of the Predictive Validity of the Suicide–Implicit Association Test. *Psychological Science*, *31*(1), 65–74. <https://doi.org/10.1177/0956797619893062>

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 New York. <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*. Chapman; 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/>

# Appendices

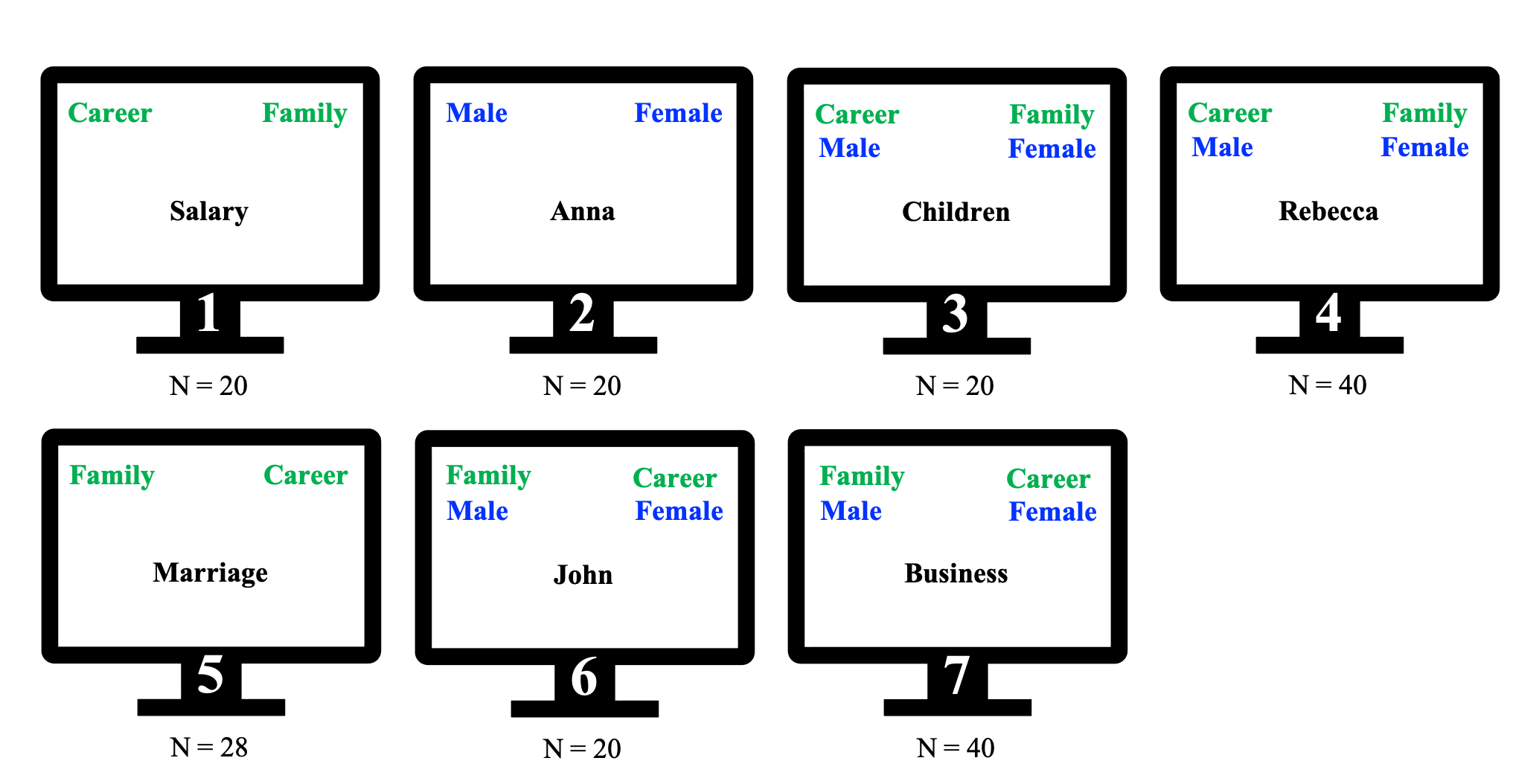
## 3.2 Appendix A

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](#ref-FAZIO1995); [Greenwald et al., 1998](#ref-GREENWALD1998)). 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](#ref-FISHER1993)). Implicit measures surpass this bias [[11]](#footnote-173) 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](#ref-FAZIO1995); [Greenwald et al., 1998](#ref-GREENWALD1998); [Hofmann et al., 2005](#ref-HOFMANN2005)). The IAT, for example, measures bias indirectly by comparing response times across a 7-block categorization task ([Greenwald et al., 1998](#ref-GREENWALD1998); [Greenwald et al., 2003](#ref-GREENWALD2003)). 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](#ref-BABCHISHIN2013)), racial biases towards Asian-Americans after terming COVID-19 the “China Virus” ([Darling-Hammond et al., 2020](#ref-DARLING-HAMMOND2020)), or measuring an individuals’ suicidal thoughts ([Tello et al., 2019](#ref-TELLO2019)).

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](#ref-HEIPHETZ2013)) are paired with attitude-categories (e.g., “Pleasant” vs. “Unpleasant,” [Greenwald et al., 1998](#ref-GREENWALD1998)) to reveal differences in association strength between two sets of categories (e.g., Heterosexual-Positive & Gay-Negative, [Steffens, 2005](#ref-STEFFENS2005)). Figure 4 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 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 4. The (in)congruent pairing of categories in the blocks 3/4 vs. 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 (). The participants’ bias score expresses the association strength between the four categories ([Nosek et al., 2005](#ref-NOSEK2005)). In the case of the GC-IAT, a positive 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](#ref-BRENDL2001)). Central to our research is the fact that both the categories and exemplars exert an effect on ([Gast & Rothermund, 2010](#ref-GAST2010)) and (in)appropriate stimuli selection directly effects the (direction of) the measured bias score.



*Figure* *4.*  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) differs 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>).

1. We replicated the search strategy from Greenwald et al. ([2020](#ref-GREENWALD2020)) on April 2022. 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. [↑](#footnote-ref-20)
2. We extended the search strategy from Greenwald et al. ([2020](#ref-GREENWALD2020)) on April 2022. 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. [↑](#footnote-ref-26)
3. Organization: <https://www.projectimplicit.net/>; Take-a-Test: <https://implicit.harvard.edu/implicit/takeatest.html> [↑](#footnote-ref-27)
4. In this study we will use the data from 15 IATs currently available via Project Implicit ([Organization]((https://www.projectimplicit.net/)); [Take-a-Test](https://implicit.harvard.edu/implicit/takeatest.html)). The following IATs are included: Age-IAT, Arab-IAT, Asian-IAT, Disability-IAT, Gender-Career-IAT, Gender-Science-IAT, Native-American-IAT, President-IAT, Race-IAT, Religion-IAT, Sexuality-IAT, Skin-IAT, Transgender-IAT, Weapons-IAT, and Weight-IAT. [↑](#footnote-ref-31)
5. We, furthermore, used the R-packages *bookdown* (Version 0.28; [Xie, 2016](#ref-R-bookdown)), *cowplot* (Version 1.1.1; [Wilke, 2020a](#ref-R-cowplot)), *dplyr* (Version 1.0.9; [Wickham et al., 2021](#ref-R-dplyr)), *foreign* (Version 0.8.82; [R Core Team, 2022](#ref-R-foreign)), *ggplot* ([Wickham, 2016](#ref-R-ggplot)), *ggtext* (Version 0.1.1; [Wilke, 2020b](#ref-R-ggtext)), *glue* (Version 1.6.2; [Hester, 2020](#ref-R-glue)), *here* (Version 1.0.1; [Müller, 2020](#ref-R-here)), *knitr* (Version 1.39; [Xie, 2022](#ref-R-knitr)), *lme4* ([Bates et al., 2015](#ref-R-lme4)), *lmerTest* ([Kuznetsova et al., 2017](#ref-R-lmerTest)), *lubridate* (Version 1.8.0; [Grolemund & Wickham, 2011](#ref-R-lubridate)), *papaja* (Version 0.1.1; [Aust & Barth, 2020](#ref-R-papaja)), *readr* (Version 2.1.2; [Wickham & Hester, 2020](#ref-R-readr)), *rmarkdown* (Version 2.14; [Allaire et al., 2022](#ref-R-rmarkdown)), *Rmisc* (Version 1.5.1; [Hope, 2013](#ref-R-Rmisc)), *scales* (Version 1.2.0; [Wickham & Seidel, 2020](#ref-R-scales)), and *tidyverse* (Version 1.3.2; [Wickham et al., 2019](#ref-R-tidyverse)). [↑](#footnote-ref-36)
6. The Ethics Review Board of *name removed for review purposes* approved the research plan on 19th of March 2021 (ref: *removed for review purposes*). [↑](#footnote-ref-38)
7. Jordan Axt, the director of Data and Methodology for Project Implicit, has confirmed on the 16th of 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 <https://osf.io/dw23y/?view_only=25b62f307a1349e7883549b473091483>. [↑](#footnote-ref-39)
8. Organization: <https://www.projectimplicit.net/>; Take-a-Test: <https://implicit.harvard.edu/implicit/takeatest.html> [↑](#footnote-ref-40)
9. Determined at the time of writing ( of September 2021) [↑](#footnote-ref-42)
10. In 2020 data was also collected for the *COVID-19* IAT. This IAT was not included because (a) the IAT was online for only a short amount of time, and (b) the IAT was included as part of a much more extensive research procedure. In addition, had we used the data from December 2020, all users from the *Asian-American* IAT would have been excluded due to missing demographic data. We thus opted to use the data from November 2020 instead. [↑](#footnote-ref-45)
11. Among others, Hofmann et al. ([2005](#ref-HOFMANN2005)) and Gawronski ([2009](#ref-GAWRONSKI2009)) provide mixed evidence accounts of implicit measures effectively reducing social desirability biases. [↑](#footnote-ref-173)