Alia Yamin ‘23
Dr. Damian Stanley
Honors College Thesis
1 May 2023

The Influence of Implicit Race Bias on Working Memory for Trait Associations

Introduction

Despite attempts towards greater racial and social equality, racial disparity remains an issue in American society (Bowdler & Harris, 2023). One source of this disparity is bias that we all hold towards different racial groups. Psychologists make a distinction between implicit bias and explicit bias (Reihl et al., 2015). Implicit biases are automatic associations that are formed outside of one's control and that can be activated without our awareness. This type of covert bias has been found to influence cognitive processes without an individual's knowledge (e.g., Stanley et al. 2011; Kubota et al. 2013; Green et al. 2007). Explicit biases are attitudes and biases that are consciously accessible and expressible (Reihl et al., 2015). Within the context of race biases, implicitly held biases are reflected in subtle microaggressions, attitudes, or beliefs that an individual may not even have perceived as being a bias (Devine 2001). However, explicit race biases are consciously and overtly racist and discriminatory behaviors, such as hate crimes.

Implicit race bias can be quantified with the Implicit Association Test (IAT-D) (Greenwald et al. 1998). In 1998, Anthony Greenwald, Debbie McGhee, and Jordan Schwartz developed a task that measured attitudes towards various groups – race, gender, and sexuality – by evaluating the degree of associations formed between different types of information (Greenwald et al. 1998). The task developed by Greenwald and his associates utilized reaction time to evaluate the degree to which an individual had associations between certain concepts or ideas. The IAT allows cognitive and behavioral scientists to develop statistical analyses of abstract cognitive processes that can occur without an individual’s awareness and have the capacity to influence other aspects of the mind.

Research into implicit racial biases has evoked many inquiries regarding the potential influence of biases on cognition, specifically face perception, trust estimation, decision making,

and memory (Brown et al. 2017; Dostch et al. 2008; Kubota et al. 2013; Stanley et al. 2011). These cognitive processes are crucial in daily life, emphasizing the importance of investigating implicit race biases.

Daily social interactions with friends, family, and coworkers required the ability to identify and analyze faces. Face perception refers to the ability to recognize faces as well as use facial expressions for certain emotional and social cues. Dr. Ron Dotsch and his colleagues investigated whether prejudices influenced the associations between an outgroup and their respective stereotypical characteristics. There were two parts to this investigation, both assessing for prejudices of the Dutch participants regarding an out-group stimulus, images of Moroccan people. Participants were first shown the same face image with different patterns of visual noise (i.e., static) superimposed each time. For each stimulus, the participants were asked to classify the face as either Chinese or Moroccan (note that the only difference between each stimulus is the pattern of noise superimposed on it). Researchers used the participants' responses to create a visual representation of each participants' internal representation of a prototypical Chinese and prototypical Moroccan face. Researchers also measured the participants' implicit biases for towards Morrocans and Chinese groups. Another group of participants were then asked to rate the prototypical faces derived from the first set of participants' responses on how criminal and trustworthy they looked. The implicit bias prejudices from the first group were correlated with the ratings of criminality and trustworthiness from the second group. When participants showed more bias against Moroccans, their internal representation of the typical Moroccan face (measured with those noisy images) was judged as less trustworthy looking and more criminal looking by other groups of participants (Dotsch et al. 2008).

The manner in which individuals gauge the trustworthiness of others has similarly been found to be shaped by implicit race biases (Stanley et al., 2011). Participants were first presented with face stimuli and asked to rate them on trustworthiness scales. Participants with a pro-White implicit race bias were more likely to rate White face stimuli as being more trustworthy than Black face stimuli (and vice-versa). This bias extended into economic decision-making as well. In a second study, participants with a pro-white bias were less likely to engage in a financially risky interaction when presented with pictures of Black partners compared to White partners. These results implied that individuals with pro-white implicit race

bias were less likely to trust Black American individuals than White American individuals (Stanley et al. 2011).

In a third study, Brown and their associates found that attention and memory were different for same- and other-race faces. There is an established social phenomenon known as the Other Race Effect (ORE) in which there is a difference in which information regarding same and out-group faces is is processed with a different attentional allocation. This suggested that individuals engaged in more in-depth processing of same- or in-group faces. A similar trend was observed in recognition memory, which refers to the ability of an individual to identify information they have encountered previously. Participants of European-American (EA) and African-American (AA) descent were given an encoding phase and a recall phase within the task. In the encoding phase, they were presented with an equal number of randomly ordered EA and AA faces while undergoing fMRI data collection. During the recall phase, participants were asked to rapidly rank a series of EA and AA face images as either 'old' (belonging to the encoding phase) or 'new' (not previously seen in the encoding phase). The results suggested that group-based face memory biases influenced the allocation of cognitive control and top-down attention during encoding, as well as showed predictive memory failure for encoding out-group faces. This suggests that race recognition memory processes are less efficient for out-group faces than in-group faces (Brown et al. 2017).

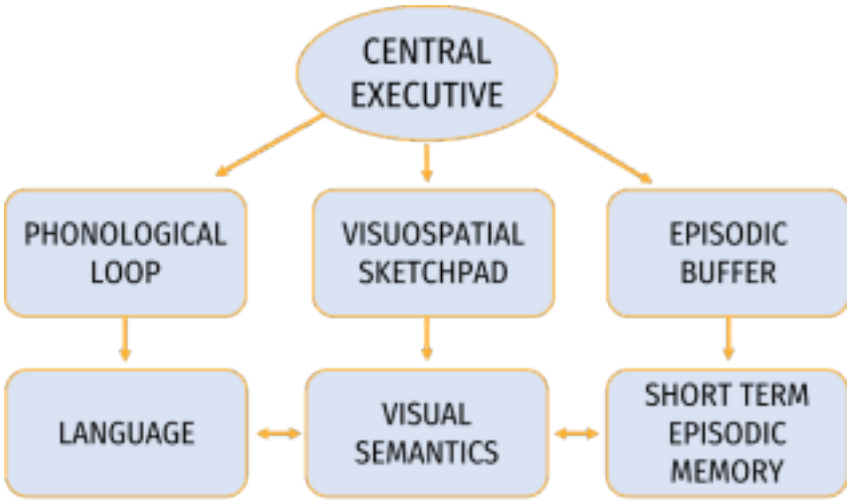
Finally, in a study focusing on decision making, participants were presented with a task modeling economic decision making during financial negotiations. In a simulated financial situation called the 'Ultimatum Game', players or participants were asked to either accept a proposed split of money and receive their portion or reject proposed splits of money and lose all chance of receiving money. This scenario was used to investigate whether participants would reject or accept financial proposals from White and Black American proposers at the same rate. When combined with the evaluation of stereotypes and implicit race bias, this was used as a means of assessing for a potential correlation between biases and objectively equal unfair financial proposals. It was found that participants with a greater pro-white implicit bias were willing to accept a greater number of offers and lower offer amounts from White American proposers than Black American proposers (Kubota et al. 2013).

From these and other studies, it has been shown that the stages of memory - encoding, storage, and retrieval – are influenced by implicit race bias. One domain that remains

unexamined is working memory. Memory can be defined as either a device in which information is held or the ability of the mind (Morris et al., 2006). Within the concept of memory, there lie three additional sub-mechanisms known as long-term memory, short-term memory, and working memory. Each of these mechanisms serves its own purpose, despite the deceptively similar terminology used to address and categorize them.

Long-term memory involves the acquisition of sensory or other types of information, the process of consolidation by which information is transformed into long-term memory, and the retrieval of long-term memory (Hawk & Abel, 2010). Long-term memory can be further divided into episodic memory and semantic memory (Estes, 2014; Hupbach et al., 2009). Episodic memory involves recalling previous experiences, including the associated details (Hupbach et al., 2009). Semantic memory refers to the process of encoding and recalling what may be considered "general knowledge" and includes the ideas, concepts, and facts learned over a lifetime (Estes, 2014). Short-term memory refers to a cognitive system that provides limited amounts of information for a limited period. There are two types of short-term memory: iconic memory and acoustic memory. Iconic memory refers to the ability to briefly mentally visualize an image after the physical stimulus is no longer present (Phillips, 2011). Acoustic memory involves the process by which memories can be briefly encoded using the associated auditory stimuli (Darwin & Baddeley, 1974).

In working memory, information is stored for a relatively short period of time; however, this differs from short-term memory in that the information is accessed and manipulated for other cognitive processes (Baddeley et al., 2020; Cowan et al., 2013; Logie et al., 2020). Once information enters working memory, it is received by the central executive (see Figure 1). This first stop within the working memory process entails processing information by directing the mind's attention to certain details and maintaining task goals. Elements of decision-making and memory retrieval are also performed by the central executive (McCabe et al., 2010). After information is processed, the phonological loop receives the auditory information, the episodic buffer integrates information and allows the brain to generate a sequence of events, and the visuo-spatial scratchpad temporarily maintains the visual and spatial stimuli (Baars & Gage, 2010; Fiez, 2016; Henry, 2010).



[Figure 1. Visual Representation of Working Memory Process]

Note. This provides a visual representation of the manner in which information interacts with working memory. Information is processed by the central executive and sorted to the appropriate sub-processes, after which it may be integrated with other items.

Working memory has been found to be influenced by a number of factors. Information stored in working memory is limited to a set amount called a memory load. On average, the working memory load that adolescents can encode is 3 to 5 chunks of information (Cowan, 2010). A greater amount of information or a more complex array of stimuli is more difficult to recall from working memory than a single simple stimulus. The familiarity of information also influences how well an individual is able to process a particular stimulus in their working memory (Brady et al., 2016). Anxiety inhibits working memory by interfering with certain relevant underlying processes (Moran, 2016).

Given previous findings on implicit bias and other aspects of cognition (see above), there is good reason to infer that implicit race bias may influence working memory as well. Implicit race bias would be expected to generally impair working memory for those an individual is biased against. This would manifest as slower reaction times and reduced accuracy when answering questions about specific stimuli. In the context of implicit race

bias, the participants' working memory performance would be worse for face stimuli of a race they were found to be biased against. A greater degree of bias would have a greater impact on the working memory performance. By expanding our knowledge of the influence of implicit race bias on various cognitive processes, we can gain a greater understanding of how biases affect our cognition in our daily lives. Working memory is used on a daily basis for a variety of tasks generally considered important, such as interviewing for jobs or providing eyewitness testimony in court. With these situations in mind, it would be important to start considering how implicit race bias may have an impact upon working memory, as well as how these processes should be quantified in an experimental setting.

To experimentally isolate and elicit working memory, we developed a delayed match-to-sample (DMS) task. Match-to-sample tasks were first developed and implemented by B. F. Skinner. They were first used to investigate an animal model of association development from working memory, which was then applied to experiments with human participants starting in the 1980s (Ferster, 1960). Miller and his associates utilized a delayed match-to-sample task to evaluate the mechanisms driving visual working memory in macaque monkeys (Miller et al., 1996). Wang and his colleagues used a DMS task in a study investigating the potential age-related changes in working memory of macaque monkeys as well (Wang et al., 2011). In 2003, the DMS was used to assess the influence of cognitive load on working memory for human participants (D'Esposito & Curtis, 2003). This type of task has proven to be reliable and valid in assessing working memory task performance for animal models and human participants.

Here, we asked whether and how implicit race bias influences working memory. We hypothesized that an individual’s ability to use information from their working memory would be influenced by implicit race bias as had been shown for other processes related to cognition and perception. That is, on average, participants would have higher rates of incorrect answers for questions in this task for faces from a racial group they are more biased against, compared to a racial group they are more biased towards. The main task, the Racial Face-Trait Association Working Memory Task (RFTA), was meant to quantify the potential relationship between working memory and implicit race bias. In addition, we collected measures of implicit and explicit race bias towards Black and White individuals, and a battery of cognitive tasks assessing perceptual similarity and recognition memory to

serve as controls.

Methods

Participants

170 participants between the ages 18 and 45 years old U.S. residents; (50% female/male; 50% Black/White) were recruited via the prolific.co platform. Other prescreening criteria included: English fluency, having lived in the US for 10 or more years, and having completed between 5 and 1000 tasks on Prolific. Note: these criteria resulted in eligible participant pools of ~700 Black participants and ~7000 White participants. After exclusions due to technical issues (n=3) and data cleaning (i.e. filtering out of poor responders, n=33), 134 participants (44.4% Black) were deemed to have data viable for analysis. IRB ethics approval (Adelphi University IRB) was obtained prior to participant recruitment and data collection. In addition, the experimental design and main hypotheses were preregistered on the Open Science Framework (https://osf.io/freut/). Participants were compensated at a rate of $10.50 per estimated hour for their time ($8.75 for an estimated 50 minutes). The median study duration was 56 minutes and 15 seconds.

Several *a priori* power analyses were conducted using G*Power platform in order to determine the minimum sample size required for an 80% power to detect a medium effect size, and the sensitivity of the sample, or the effect size value of the total number of participants at 80% power (Faul et al., 2007). To achieve an 80% power for a medium effect size (cohen's d = 0.5) at α = .05, a sample size of N = 34 was required for a single sample t test. A sample size of N = 150 at 80% would have the sensitivity to detect an effect size (cohen's d) of 0.230. For a point biserial correlation, an 80% power for a medium effect size at α = .05 would require a sample size of N = 82. There is sufficient sensitivity to detect effect sizes of 0.224 and above at 80% power with 150 participants.

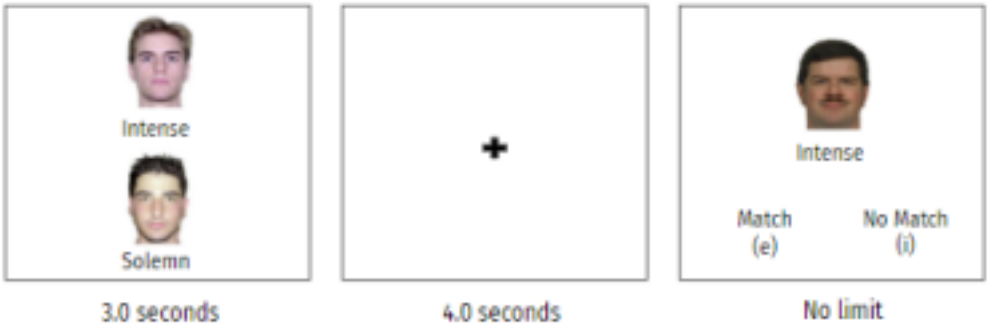
Materials/Apparatus

The participants that took part in this study accessed it from the platform prolific.co and completed the tasks on their laptops or computers. The surveys were administered using the Qualtrics survey platform (Provo, Utah), the tasks were developed using

jsPsych (de Leeuw, 2015) and were hosted on Pavlovia.org. Face stimuli were obtained from the Eberhardt Laboratory Face Database (the Race Face-Trait Working Memory and Race Recognition Memory tasks) and the Hughes laboratory (Race Perceptual Discrimination Task; Hughes et al, 2019).

Tasks

Race Face-Trait Working Memory Task



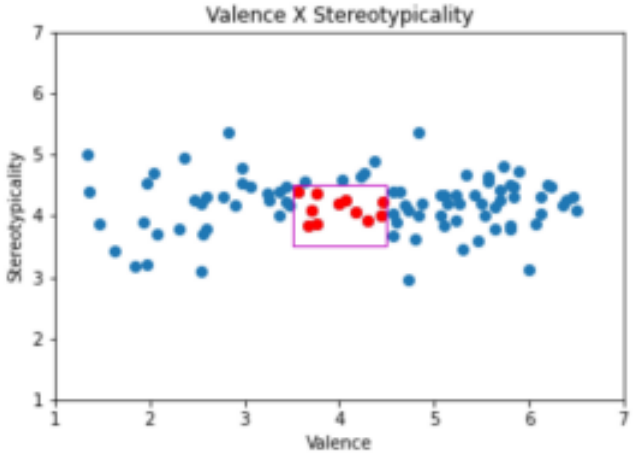
[Figure 2. Sample Sequence from Race Face-Trait Working Memory Task]

Note. Participants would first see the face array with the two face-trait pairs, followed by the 4 second long focus cross, and the 'no match'-'match' question. The assessment was followed by the Race Face-Trait Working Memory task.

The Race Face-Trait Association Working Memory Task (rFW-WMT) was adapted from the delayed match-to-sample paradigm (Blough, 1959). For each trial, the participants were first shown a sample array of faces (array size = 2-4; 'Sample' period) each of which was paired with a neutral (with respect to valence and Black/White stereotypicality) word describing a trait (e.g. 'Obvious'), and displayed for 2.5 to 5 seconds (depending on array size; 1.25 secs per face-trait pair). Then the participant was shown a fixation cross for 4.0 seconds ('Delay' period). This was followed by a 'Match' period, in which the participant was presented with a single face-word pair and asked to indicate whether the exact pair was a 'Match' (press 'e') to one of the pairs in the sample array or whether there was 'No Match' (press 'i'); participants had unlimited time to respond (see Figure 2). Importantly, non-matches did not contain any novel stimuli (all were present in the sample period), only novel pairings.

Face sample arrays were composed of all White or all Black individuals. Trait words were selected from a list of 638 positive/neutral/negative personality traits (*638 Primary Personality Traits*) and the 6 most neutral and non-stereotypical (with regards to Black/White racial stereotypes) identified in a previous study (see below). In total, participants viewed 96 trials (48 per target race, 32 per Array Size) in two blocks of 48 trials per block. Participants were also provided with 6 practice trials (3 Black, 3 White) to familiarize them with the task. The independent variables were condition (Race, set size, match type) and the dependent variables are accuracy and reaction time (for correct responses only).

Selecting Neutral and Non-Stereotypical Traits



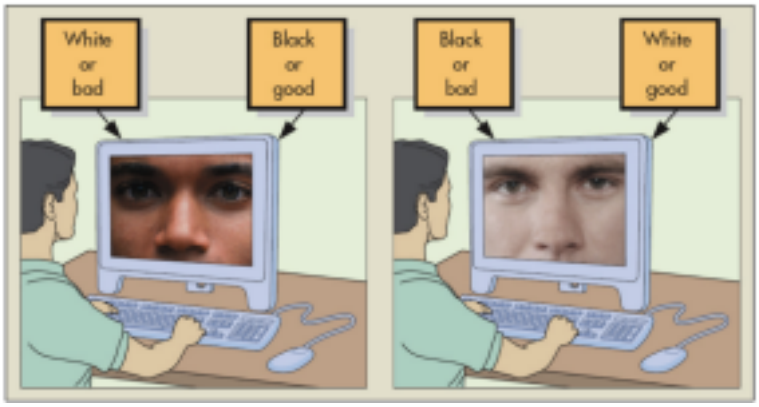
[Figure 3. Determining the Most Neutral and Non-Stereotypical Traits]

Note. This is a visual representation of the process by which the 6 traits used in rFW-WMT were chosen from a list of 638 total traits.

This preliminary task was used to identify trait words that were neutral with respect to

valence and Black/White stereotypicality for use in the rFW-WMT (see Figure 3). Participants (recruited via Prolific, N=??) were asked to rate the Black/White stereotypicality – on a scale of 1 (‘Strongly associated with Black Americans’) to 7 (‘Strongly associated with White Americans’) – and valence – on a scale of 1 (‘extremely negative’) to 7 (‘extremely positive’) – of 96 personality traits (32 each of Positive, Neutral, and Negative valence; taken from a list of 638 personality traits (see Massachusetts Institute of Technology, n.d.).

Implicit Association Test



[Figure 4. Implicit Association Test for Implicit Biases]

Note. Participants completed the IAT in two blocks and were given images similar to the ones displayed on the monitors above.

The Implicit Association Test (IAT) is a measure of the strength of the association between concepts (Greenwald et al., 1998). Participants were shown 200 trials (including single-category and dual-category types). There were a total of 120 dual category trials of

interest in which participants sorted words into Good and Bad categories, and faces into Black and White categories, simultaneously (see Figure 4). Two blocks (20 and 40 trials respectively) had a ‘congruent’ mapping (Good with White, Black with Bad), and two other blocks (20 and 40 trials respectively) an ‘incongruent’ mapping (Good with Black, White with Bad) ([in]congruency is defined with regard to stereotype norms in the U.S.). The resulting standard measure (calculated according to the algorithm in Lane et al., 2007) was called the “IAT D score” which ranges from -2 to 2, with positive scores indicating pro-white bias.

Perceptual Discrimination Task

[Figure 5. Perceptual Discrimination Task]

Note. The row of faces represents the pool of morphs from which the 50% and other percent morph faces were selected for the task.

The Perceptual Discrimination Task (PDT) evaluated how sensitive participants are to variation in face similarity across different races. Stimuli consisted of 8 Black and 8 White faces that were used to make 4 Black and 4 White face morph continua (in morph increments of 10%) (Hughes et al 2019). Each trial began with a fixation cross (duration=350ms). Then the

participants were shown two faces (duration=500ms) from the same morph continua; the 50% morph (32 trials) and another morph that is either identical (32 trials) or different (ranging from 10% to 50% different; 80 trials). After 500ms, the faces were replaced by a fixation cross, and participants indicated via key press whether the 2 faces were identical or different (see Figure 5). Participants completed a total 112 trials. The independent variables were condition (black/white faces, degree of morph), and the dependent variables were accuracy and reaction time (for correct responses only).

Race Recognition Memory Task

[Figure 6. Race Recognition Memory Task]

Note. The encoding phase of this task is demonstrated with the top row of images and the recall portion of the task is depicted by the bottom row of images with the ‘x80’.

The Race Recognition Memory Task (rRMT) functioned as a means of evaluating the participants’ recognition memory and whether it varied between races. In the recognition memory task, 80 faces (40 White, 40 Black) were used as stimuli (Hughes et al 2019). In the encoding phase, participants were shown a sequence of 40 faces (20 White, 20 Black; 2secs per face) selected at random from the pool of 80. Participants were instructed to memorize the faces, as they will answer questions about them later. Following the encoding phase, participants completed a 3 minute distractor task in which they identified differences between 2 images. participants then proceeded to the test phase, in which they were presented with each

of the 80 faces in random order. Participants indicated via key press whether each target is ‘Old’ (i.e., presented during the encoding phase) or ‘New’ (see Figure 6). The independent variables were condition (black/white faces) and the dependent variables were accuracy and reaction time (for correct responses only).

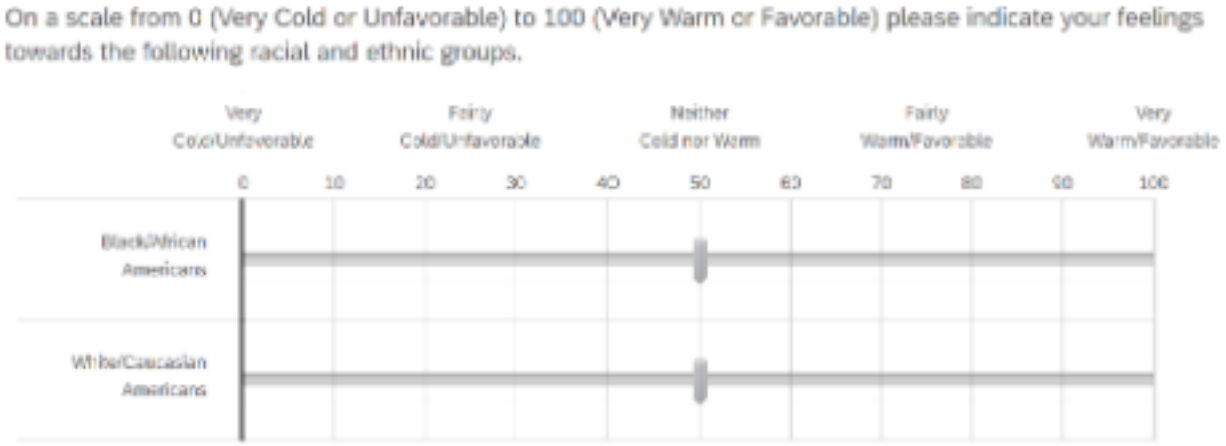
Internal/External Motivation to Respond Without Prejudice Scales

[Figure 7. Internal/External Motivation to Respond Without Prejudice Scales]

Note. Shown above are two of the questions given to participants completing the IMS-EMS portion of the study.

The Internal/External Motivation to Respond Without Prejudice Scales (IMS-EMS) (Plant & Devine, 1998) consisted of prompts focusing on the extent to which individuals are internally or externally motivated to not appear prejudiced towards ‘Black people’. The participants indicated the degree to which they agreed with 10 statements on a nine point scale that ranged from ‘Strongly Disagree’ to ‘Strongly Agree’ (see Figure 7).

Race Feelings Thermometers



[Figure 8. Race Feelings Thermometers]

Note. Participants were asked to complete the same questionnaire as shown above.

The race feelings thermometer (Axt, 2018) asked the participants to indicate on a scale from 0 ('Very Cold or Unfavorable') to 100 ('Very Warm and Favorable') their feeling towards Black/African Americans and White/Caucasian Americans (see Figure 8).

Protocol

Following recruitment on Prolific, participants provided informed consent and indicated their commitment to completing the task. Participants then answered a set of demographic questions, following which, they were redirected to a page with the rFW-WMT. They received instructions for the task, completed a set of practice trials, and then began the working memory task. TFollowing the working memory task, participants complete the Black/White Pleasant/Unpleasant IAT. Next, they completed the PDT and the two-section rRMT (counterbalanced order). Upon completion of the tasks, participants received a completion code that they entered in the original qualtrics survey. Then, they completed the explicit race measures (Race Feelings Thermometer, IMS/EMS, and contact measures). Finally, the

participants viewed the debriefing (see Addendum). Participants were then redirected back to to receive their compensation for successfully completing the task.

Data Analysis

Data Cleaning

Participants who exhibited evidence of lack of engagement with the implicit measures of bias (IAT) will be excluded from the analyses featuring IAT measures. Lack of engagement was established with unusually fast (<300ms) or slow (>10,000ms) reaction times as well as unusually repetitive response behavior or an unusually high error rate (IAT; see Lane et al, 2007 and Stanley et al, 2011).

In addition to examining the relationships between individual survey measures of explicit bias (Contact Measures, MRS, SRS, IMS/EMS) and our primary variables of interest, we used component analysis (e.g. factor analysis or ICA) on survey measures of explicit attitudes (Contact Measures, MRS, SRS, IMS/EMS) to reduce the inevitable redundancy between these measures, for use as covariates in other analyses.

Target enrollment for this project was 240 participants (with an expected 15-20% attrition rate). However, for the purposes of this thesis, only the data from the initial 170 participants were analyzed (data collection ongoing). Following data preprocessing and cleaning, the data of 36 participants were excluded, resulting in a final sample size of 134 participants.

Analysis

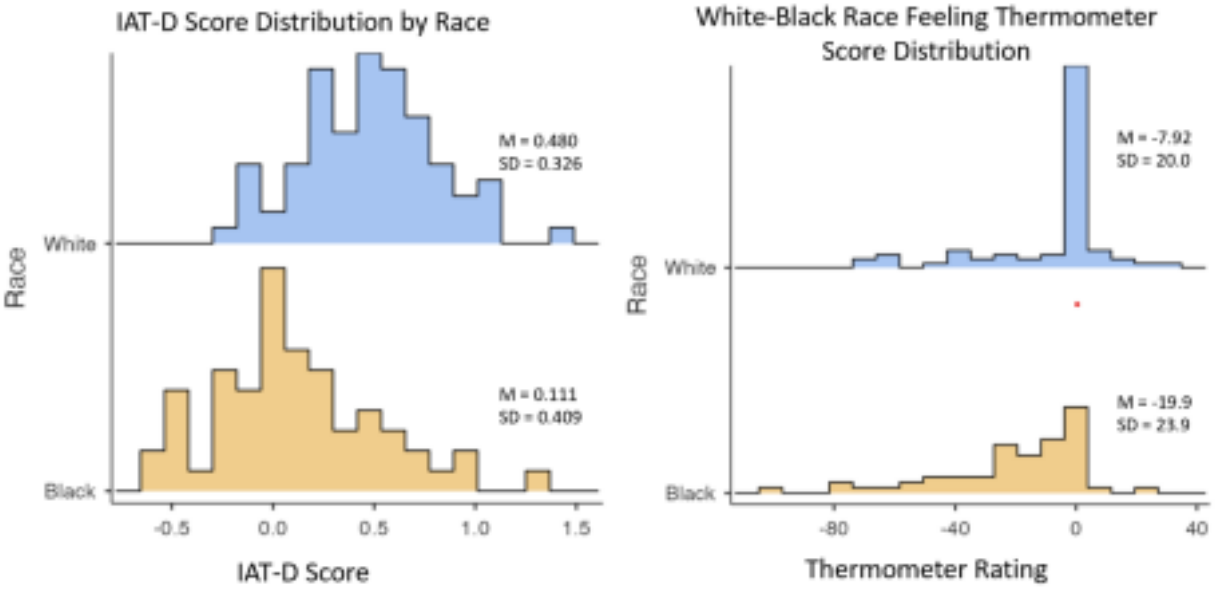
Statistical analysis consisted of standard statistical tests, including: Student's t tests, bootstrapped confidence intervals, Pearson's correlation, one- and two-way repeated measures ANOVAs, multiple linear regression, and the calculation of signal detection theory parameters d-prime and criterion. Because these are standard statistical tests they will not be further described here.

Results

IAT

On average, the participants had an overall pro-white bias (N = 134, M = 0.317, SD=0.407). The Black American participants had, on average, a slight pro-white bias (N = 59, M = 0.111, SD=0.409). The White American participants had, on average, a pro-white bias as well (N = 75, M = 0.480, SD=0.326) (see Figure 9). There was a statistically significant difference in mean IAT-D score in the two groups, White American participants and Black American participants ($t(132) = 5.81$, $p < 0.01$).

With regards to evaluating ingroup/outgroup prejudices through IAT-D scores, participants rated ingroup members as more warm/favorable than outgroup members (N=134, M=0.220, SD=0.468). Black participants had a pro-outgroup bias (the value of their pro-White bias above multiplied by -1), whereas White participants had a pro-ingroup bias (reflected in their pro-White bias mentioned above).



[Figure 9. Distribution of IAT-D Score and White-Black Race Feeling Thermometer Scores by Race]

Note. Distributions of IAT-D scores and White-Black race feeling thermometer ratings by participant race.

White-Black Race Feeling Thermometers

On average, participants rated Black Americans more warmly/favorably than White Americans (N = 135, M = -13.3, SD = 22.5). Black participants had, on average, a higher pro-black bias in thermometer ratings (N = 60, M = -19.9, SD = 23.9) compared to White participants (N = 75, M = -7.92, SD = 20.0) (see Figure 10). There was a statistically significant correlation between overall race feeling thermometer ratings and IAT-D ($r(135) = 0.291$, $p < 0.001$).

With regards to ingroup/outgroup biases in thermometer ratings, participants rated ingroup members as more warm/favorable than outgroup members (N=135, M=4.46, SD=25.8). Black participants had a pro-ingroup bias (the value of their pro-White bias above multiplied by -1), whereas White participants had a pro-outgroup bias (reflected in their pro-Black bias mentioned above).

Working Memory

Overall Performance

The participants (N = 133) had an overall average of 0.705 accuracy (SD = 3.45) and 7.30 second reaction time (SD = 6.139) on the working memory task. On average, Black American participants (N = 59) had a 0.699 accuracy (SD = 0.129) and 7.286 second reaction time (SD = 7.27). On average, White American participants (N = 74) had a 0.710 accuracy (SD = 0.125) and 7.316 second reaction time (SD = 7.27). The difference in mean overall reaction time grouped by race was not statistically significant ($t(131) = 0.523$, $p = 0.602$). The difference in the mean accuracy group by race was not statistically significant ($t(131) = 0.724$, $p = 0.471$).

There was a statistically significant difference in the participants' average working memory performance across array size ($F(2, 264) = 77.0$, $p < 0.001$). The difference in mean accuracy for questions referring to an array sizes of 2 and 3 was statistically significant ($t(132) = 6.66$, $p < 0.001$). The difference in mean accuracy for array sizes 2 and 4 was statistically significant ($t(132) = 11.12$, $p < 0.001$). The difference in mean accuracy for array sizes 3 and 4 was statistically significant ($t(132) = 6.61$, $p < 0.001$).

The difference in average reaction time across array size was statistically significant ($F(2, 264) = 11.8$, $p < 0.001$). The difference in mean reaction time for questions referring to an array sizes of 2 and 3 was statistically significant ($t = -4.0149$, $p < 0.001$). The difference in

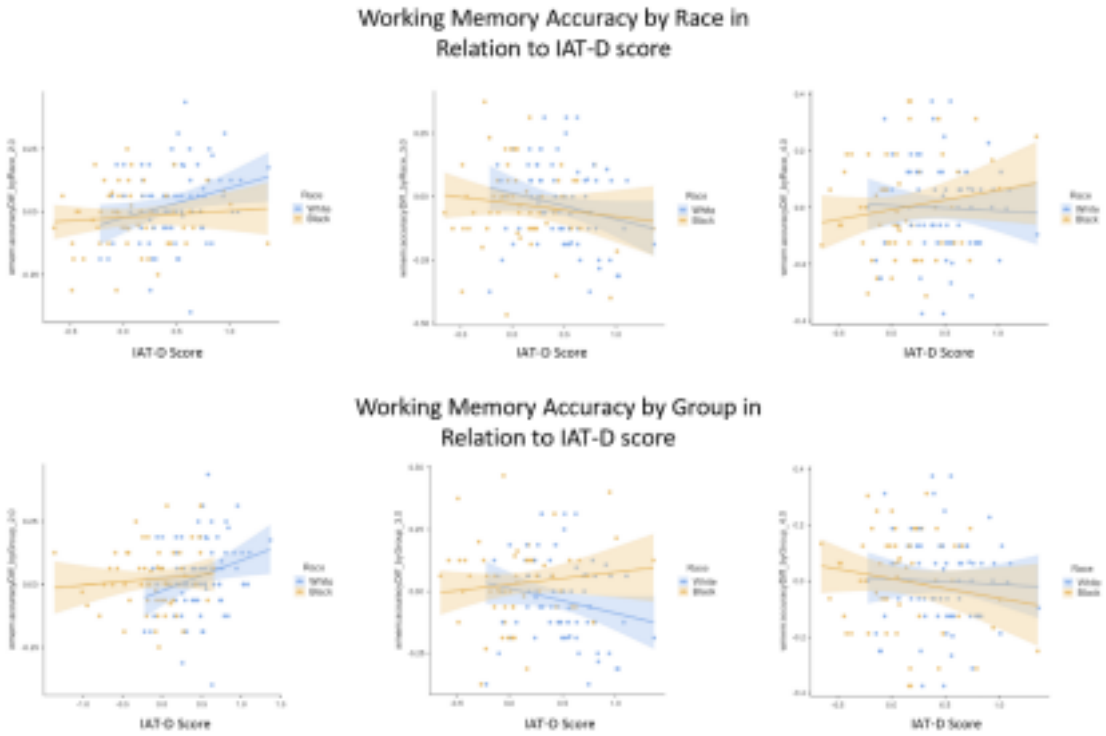
mean reaction time for questions referring to an array sizes of 2 and 4 was statistically significant ($t = -3.6022$, $p < 0.001$). The difference in mean reaction time for questions referring to an array sizes of 3 and 4 was not statistically significant ($t = 0.0363$, $p < 0.999$).

The Influence of Race on Working Memory Performance

To investigate the effect of target race on accuracy we ran a 2-way repeated measures anova (Race x Array Size). This identified a main effect of Array Size ($F(2, 260) = 44.5162$, $p < 0.001$), but no effect of Race ($F(1, 130) = 1.9554$, $p = 0.164$) and no interaction ($F(2, 260) = 0.0738$, $p = 0.929$). To further explore whether implicit race bias influenced performance, we added participant's IAT scores to the previous 2-way anova as a between participant covariate. This analysis identified a statistically significant 3-way interaction between Race, Array Size, and IAT-D Score ($F(2, 260) = 4.7310$, $p = 0.010$).

We performed the same analysis with log-transformed reaction times as the dependent variable and again there was a main effect of Array Size ($F(2, 260) = 7.465$, $p < 0.001$), and no statistically significant interaction between Race, Array Size and IAT-D scores ($F(2, 260) = 0.223$, $p = 0.801$). There was no statistically significant interaction between array size and IAT-D performance for reaction time ($F(2, 260) = 0.315$, $p = 0.730$).

To further understand what was driving the interaction between Race, Array Size, and IAT D-Score, we calculated the difference in accuracy (white-black) for each Array size and correlated the result with the IAT D-Scores. Participants' IAT D-Scores were positively correlated with the Race-related difference in accuracy for Array Size 2 ($r(130) = 0.244$, $p = 0.005$), but not for Array Sizes 3 ($r(130) = -0.151$, $p = 0.084$) and 4 ($r(130) = 0.056$, $p = 0.527$).



[Working Memory Accuracy in Relation to IAT-D]

Note. The difference in working memory accuracy by race and group were examined at each array size (2, 3, 4) in comparison to IAT-D score.

To determine whether this was independent of participant's explicit biases we ran a multiple regression with IAT D-Scores and White-Black thermometer ratings as predictors and Race-related difference in accuracy for Array Size 2 as dependent variable (R^2 = 0.0717, t(130) = -1.69, p = 0.094). This analysis showed that IAT D-Scores predicted Race-related accuracy differences at Array Size 2 (p = 0.002) independently of participants' explicit race biases (p = 0.198).

The Influence of Group Membership on Working Memory Performance

To investigate the effect of target race on accuracy we ran a 2-way repeated measures anova (Group x Array Size). This identified a main effect of Array Size (F(2, 260) = 65.597, p <0.001), but no effect of Group (F(1, 130) = 0.994, p = 0.321) and no interaction (F(2, 260) =0.608, p = 0.545).

To further explore whether implicit race bias influenced performance, we added participant's IAT scores to the previous 2-way anova as a between participant covariate. This analysis identified a statistically significant 3-way interaction between Group, Array Size, and IAT-D Score ($F(2, 260) = 6.416$, $p = 0.002$). We performed the same analysis with log-transformed reaction times as the dependent variable and again there was a main effect of Array Size ($F(2, 260) = 6.203$, $p = 0.002$), and no statistically significant interaction between Group, Array Size and IAT D-Scores ($F(2, 260) = 0.995$, $p = 0.371$). There was no statistically significant interaction between array size and IAT-D performance for reaction time ($F(2, 260) = 2.943$, $p = 0.054$).

To further understand what was driving the interaction between Group, Array Size, and IAT D-Score, we calculated the difference in accuracy (ingroup-outgroup) for each Array size and correlated the result with the IAT D-Scores. Participants' IAT D-Scores was positively correlated with the Group-related difference in accuracy for Array Size 2 ($r(131) = 0.168$, $p = 0.055$) and Array Size 4 ($r(131) = 0.045$, $p = 0.608$), but not for Array Size 3 ($r(131) = -0.257$, $p = 0.003$). The correlation between Array Size 3 and IAT-D scores was statistically significant.

To determine whether this was independent of participant's explicit biases we ran a multiple regression with IAT D-Scores and White-Black thermometer ratings as predictors and Group-related difference in accuracy for Array Size 2 as dependent variable ($R^2 = 0.0321$, $t(131) = 1.752$, $p = 0.173$). This analysis showed that IAT D-Scores predicted Group-related accuracy differences at Array Size 2 ($p = 0.082$) independently of participants' explicit race biases ($p = 0.468$). We also ran a multiple regression with IAT D-Scores and White-Black thermometer

ratings as predictors and Group-related difference in accuracy for Array Size 3 as dependent variable ($R^2 = 0.066$, $t(131) = 0.855$, $p = 0.394$). This analysis showed that IAT D-Scores predicted Race-related accuracy differences at Array Size 3 ($p = 0.004$) independently of participants' explicit race biases ($p = 0.753$).

Discussion

We set out to investigate the relationship between implicit race bias and working memory for face-trait associations. To achieve this we analyzed participants' working memory

performance on a delayed match-to-sample task (rFW-WMT) alongside their IAT-D score. In support of our main hypothesis, we found that while neither race nor group membership predicted any differences in working memory performance, there was an interaction between race, working memory load (i.e. array size), and implicit race bias. Namely, IAT D scores predicted race-related differences in working memory performance at the lowest load (array size 2) but not for higher loads (array sizes 3/4). Importantly, this effect was independent of explicit race bias.

Although there were no race-related differences in working memory performance, variation in working memory performance was significantly correlated with IAT-D scores. These results align with the results of Phelps et al. (2000). In a study that examined responses in the amygdala to black and white faces, they did not see a statistically significant difference in amygdala responses between black and white faces until the data was correlated with IAT-D scores (Phelps et al. 2000).

It is interesting that, while we originally predicted an overall influence of implicit race bias, it was only seen at the lowest level of working memory load. One possible explanation for this is that the task was very difficult for participants, which may have suppressed any implicit bias-related variation at higher working memory loads. There are a number of ways that the difficulty of the task could be reduced.

As mentioned in the methods section, the rFW-WMT was very long, with 96 trials long split into two blocks. The sheer length of the paradigm may cause participants to feel mentally fatigued. There is evidence of a relationship between fatigue and working memory performance (Roy et al. 2013; Westbrook et al. 2018). The exhaustion associated with prolonged, mentally intensive tasks can be alleviated by incorporating breaks into the task. The rFW-WMT could be broken into more blocks in which participants would be given a different, less intensive task or a blank screen for a short period of time between trials. This would be helpful in reducing mental fatigue.

To account for mental fatigue in the collected data, there should be an analysis conducted on participant performance on earlier blocks in comparison to later blocks. Should there be a significant difference between performance on the earlier block than on the later block, this would indicate that mental fatigue may be playing a role in affecting working memory performance and should be accounted for in the study. Then the later block would not

be included in the data used for the study, as there would be an outside influence affecting the participants' performance.

Another factor in the task difficulty was the number of items the participants are required to remember. Throughout the task, participants are asked to remember 4 to 8 items (depending on face array size). Array Sizes 3 and 4 exceeded the average working memory load. This may be the underlying reason why there was a lack of working performance difference for the larger array sizes. Participants may have had difficulty with the time limits as well. We could increase

the amount of time per face-trait from 1.25 seconds to 1.5 or 2 seconds. This would give the participants a greater amount of time to break previous associations and form new associations between faces and their paired traits. Increasing the amount of time may allow participants to overcome the limit of working memory loads.

With only 6 faces and 6 traits in the arrays, previous associations between particular items may have increased the difficulty of the task. The limited number of faces and traits would result in participants seeing the same faces and traits sequentially. After forming an association between a particular trait and face, participants would frequently have to dissolve those associations and form new associations in a short period of time. To avoid the interference of these associations altogether, it would be prudent to increase the pool of faces and traits used in the task. This would reduce the likelihood of face and traits repeating in sequential trials.

It would be important to consider the influence of confounding cognitive processes on working memory performance. Perceptual similarity and recognition memory may have significant influence on working memory performance. The influence of these processes can be regressed out in future analyses using the data from the Perceptual Discrimination Task (PDT) and Race Recognition Memory Task (rRMT). Controlling for these variables may change the statistical significance of working memory performance on the rFW-WMT.

We found that implicit race bias may indeed influence working memory performance when associating traits with Black compared to White faces. This may have important implications for a variety of social domains–including education, employment, healthcare, and civilian-law enforcement interaction. Despite overarching democratic ideals and egalitarian goals, American society is rife with inequality. Compared to White Americans, Black

Americans would be >3 times as likely to be killed by police (Schwartz & Jahn, 2020), >2x as likely to be

unemployed (Williams & Wilson, 2019), and suffer worse health and healthcare outcomes (Riley, 2012). There is also evidence of great disparities in wrongful convictions with Black Americans being seven times as likely to be falsely convicted of crimes (Gross et al., 2022 p. 8). This is due in part to overt individual racism and structural biases that discriminate against Black Americans, however, these ideologies and institutions are not entirely responsible for the rampant discrimination. Implicit biases may be influencing our social behaviors despite our explicit intentions (Stanley et al, 2011; Kubota et al., 2013). Understanding specifically how implicit biases shape our behaviors is critical to mitigating their impact throughout society.

Understanding how implicit biases play a role in such interactions is critical for mitigating their impact. Working memory is used to interpret recent experiences and commit them to memory (Pollmann & Schneider, 2022). If there is a discrepancy in how well an interviewer is remembering the information or experience of interviewing potential candidates for employment this may cause a bias in the decision-making process. Future research should replicate this finding and examine how it may impact consequential behaviors and occur in real world interactions.

.

Debriefing for Measures of Implicit Attitudes

Psychologists understand that people may not say what's on their minds either because they are unwilling or because they are unable to do so. For example, if asked "How much do you smoke?" a smoker who smokes 4 packs a day may purposely report smoking only 2 packs a day because they are embarrassed to admit the correct number. Or, the smoker may simply not answer the question, regarding it as a private matter (These are examples of being unwilling to report a known answer). But it is also possible that a smoker who smokes 4 packs a day may report smoking only 2 packs because they honestly believe

they only smoke about 2 packs a day (Unknowingly giving an incorrect answer is sometimes called self-deception; this illustrates being unable to give the true answer).

The unwilling-unable distinction is like the difference between purposely hiding something from others and unconsciously hiding something from yourself. The tests for implicit attitudes that you completed today make it possible to penetrate both of these types of hiding. These tests measure implicit attitudes and beliefs that people are either unwilling or unable to report.

Origins of Implicit Attitude Measurement

Different measures of implicit attitudes were originally developed as devices for exploring the unconscious roots of thinking and feeling, but can also serve as a tool to gain greater awareness about one's own unconscious preferences and beliefs.

Many years ago, Fyodor Dostoyevsky wrote: "Every [person] has reminiscences which [they] would not tell to everyone but only [their] friends. [They have] other matters in [their] mind which [they] would not reveal even to [their] friends, but only to [themselves], and that in secret. But there are other things which a [person] is afraid to tell even to [themselves], and every decent [person] has a number of such things stored away in [their] mind."

These lines from Dostoyevsky capture two concepts that these measures of implicit attitudes help us examine. First, we might not always be willing to share our private attitudes with others. Second, we may not be aware of some of our own attitudes. Your results on these tests of implicit attitudes may include both components of control and awareness.

Understanding and Interpreting Implicit Attitude Results

Although these measures of implicit attitudes were developed for research use, they have clear potential for application outside the laboratory. The results may be especially interesting if you find that they reveal an automatic association that you could not control. For example, you may believe that women and men should be equally associated with "science" - yet, your automatic associations may show that you (like many others) associate male (more than female) with science.

How might you use experiences with these various tests to think about the

implications of unconscious thoughts and feelings? We can tell you about the types of questions we considered after taking a test measuring implicit age attitudes: "What does it mean that we show an automatic association between old and unpleasant? What is the source of such knowledge? Should we be disturbed by the fact that we possess such associations? If we are (and indeed we are!), what might we do about it?"

We urge caution in using these implicit attitude tests to reach conclusions about yourself or others. You might wonder, for example, if these tests can be used to make decisions about yourself (e.g., what should I buy, where should I go to school, etc.) If you are female, and you show a greater association between male and science (as the majority of men and women do), should you decide to avoid a scientific career? Our opinion is: Most definitely not! This test result might instead prompt you to take note of the broad reach of gender stereotypes and to ask what it means to be setting out towards a scientific career in a world in which so many people automatically associate science with male (including perhaps yourself).

Can (or should) people use this test to make decisions about others? Can one, for example, use this test to measure somebody else's automatic racial preference, and use it to decide that they should or should not serve on a jury? We assert that these implicit attitude tests should not be used in any such way. Especially at this early stage of their development, it is much preferable to use it mainly to develop awareness of one's own and others' automatic preferences and stereotypes. Using implicit attitude tests as the basis for making significant decisions about self or others could lead to undesired and unjustified consequences.

We hope you have been able to take something of value from the experience of taking one or more of these tests.

Ethical Considerations

You are likely already aware that tests of implicit attitudes have the potential to reveal troubling aspects of 'human nurture.' They therefore pose significant possibilities for misuse. If you are considering use of an implicit attitude test in research, this source of concern will of course be taken into account in developing your research plans, in accordance with

safeguards present in the institutional setting of your research.

As we mentioned before, measuring implicit attitudes has clear potential for application outside the research laboratory. The following possible misuses might arise when operating outside the laboratory (and therefore operating beyond the scope of safeguards present at research institutions). First, people may use a measure of implicit attitudes to make decisions about themselves: what should I buy, where should I go to school, etc. These seem, at least on the surface, to be acceptable (I may use any method I like, including looking at tea leaves, to decide that I want to work at Firm X, Y or Z rather than at Firm A, B or C). Second, people may use them to make decisions about others: for example, one use might be to ensure that people showing a certain degree of automatic racial preference cannot participate in decision-making in racially sensitive situations. Third, people may use them to investigate others' likes and dislikes, causing the individuals harm in the process. We, along with the investigators who have been involved in developing different measures of implicit attitudes, urge careful consideration of costs associated with these possible misuses in contemplating any application outside of the laboratory.

Thank you again for your participation. For more information, please ask the researcher who administered your study or see the website below (from which the text was adapted for this debriefing form). https://implicit.harvard.edu/implicit/demo/

In this study, we are interested in people’s natural behavior when they are in a social context, and so it was important to create a situation that was realistic. For this reason, we could not tell participants all the details of the study until afterward. If we did, past research has shown that participants try to regulate their feelings and counteract any effects they may have on the experimental measures. This mild type of deception is often necessary for us to be able to learn about how people act in real-life social situations. Please keep in mind that all of your responses are completely anonymous and will never be attached to your name.

Addendum B

1. Have any data been collected for this study already? (optional)

No

2. What's the main question being asked or hypothesis being tested in this study? (optional)

Despite overarching democratic ideals and egalitarian goals, American society is rife with inequality. For example, compared to White Americans, Black Americans are >3 times as likely to be killed by police (Schwartz & Jahn, 2020), >2x as likely to be unemployed (Williams & Wilson, 2019), and suffer worse health and healthcare outcomes (Riley, 2012). This is due in part to overt individual racism and structural biases that discriminate against Black Americans, however, that is not the whole picture. In addition, implicit biases – which are automatic, difficult to control, and can be activated without our awareness – influence our social behaviors despite our explicit intentions (Stanley et al, 2011; Kubota et al., 2013). Understanding specifically how implicit biases shape our behaviors is critical to mitigating their impact throughout society.

Research into implicit race bias has revealed its influence over various cognitive processes including face perception, trust estimation, social decision-making, and recognition memory (Brown et al. 2017; Dostch et al. 2008; Kubota et al. 2013; Stanley et al. 2011). However, there have been little-to-no studies that have explored how implicit bias may impact working memory, a sub-mechanism of memory in which information is stored for a relatively short period of time, and can be accessed and manipulated for other cognitive processes (Cowan, 2013; Baddeley et al., 2021).

To address this gap in the literature, we will investigate the influence of black/white implicit race bias on working memory for face-trait associations in a novel delayed-match-to-sample paradigm. To control for potential perceptual discrimination and recognition memory confounds, we will also assess those variables using previously established tasks (Hughes et al, 2019). Participants will first complete the race working memory task

(RWMT), followed by a African-American/European-American+Good/Bad implicit association test (IAT, Greenwald et al., 1998). Following these main tasks, participants will complete a Black/White face recognition memory task (bwFRMT) and a Black/White face perceptual discrimination task (bwFPD; in counterbalanced order), and finally surveys assessing explicit race attitudes.

Our main research question (Is there a relationship between working memory for face-trait associations and implicit race bias?) is exploratory and therefore we do not have strong confirmatory hypotheses as to the nature of the relationship. However, we do have

beliefs about the direction of effects for the RWMT, which we detail here. In addition, we have confirmatory and exploratory hypotheses regarding the other tasks in the dataset (the IAT, bwFPD, and bwFRMT.

Specific Hypotheses (confirmatory hypotheses are marked with a C in parentheses before the hypothesis, exploratory with an E):

1. Race Working Memory Task (RWMT)

a. (C) Participants' overall working memory performance will be above chance (i.e., 50%).

b. (C) Task Performance: Larger face array sizes will be associated with poorer working memory performance (i.e., longer reaction times and less accurate responses).

c. (C) Performance analyses ingroup/outgroup: Working memory performance (collapsed across face array size) will be higher for ingroup, as compared to outgroup.

d. (E) Performance analyses implicit bias: Participants' pro-white (i.e., white-black) IAT scores will have a negative relationship with differences in response reaction times for white minus black faces (collapsed across face array size).

e. (E) Participants' pro-white (i.e., white-black) IAT scores will have a positive relationship with differences in accuracy for white minus black faces (collapsed across face array size).

f. (E) As working memory load increases (i.e., face array size increases) the magnitude of ingroup/outgroup and race-attitude effects will increase.

g. (E) Exploratory signal detection theory analyses will be conducted to determine whether any observed effects are reflected in variation in sensitivity or criterion threshold.

h. (E) The influence of implicit race biases on performance metrics and signal detection parameters will be independent of the influence of explicit biases. 2. Implicit Association Test (IAT)

a. (C) As has been found in many previous studies, Black participants will, on average, have relatively unbiased average IAT scores (i.e., mean ≅ 0), and White participants will, on average, have pro-white IAT scores (i.e., mean > 0.0). 3. Black/White Face Perceptual Discrimination Task (bwFPDT)

a. (C) Performance Metrics (i.e., Accuracy and RT) will worsen (i.e., decrease and

increase, respectively) as the morph distance between target and comparison morph decreases.

b. (C) Performance analyses ingroup/outgroup: Participants will exhibit decreased reaction times and increased accuracy for ingroup as compared to outgroup faces. c. (E) Performance analyses implicit bias: Participants' IAT pro-white (i.e., white-black) IAT scores will have a negative relationship with differences in response reaction times for white minus black faces. Participants' IAT pro-white (i.e., white-black) IAT scores will have a positive relationship with differences in accuracy for white minus black faces.

d. (C) Signal Detection analyses ingroup/outgroup: Participants will show no difference in overall sensitivity between outgroup and ingroup faces, but will show decreased criterion thresholds for outgroup compared to ingroup faces.

e. (E) Signal Detection analyses implicit bias: Participants' pro-white (i.e., white-black) IAT scores will have a positive relationship with differences in criterion thresholds for white-black faces. There will be no relationship between participants' IAT scores and overall sensitivity.

f. (E) Psychometric curves describing accuracy as a function of distance between target and comparison morph will have a higher thresholds for outgroup compared to ingroup faces.

g. (E) Race-related threshold differences (White minus Black) for psychometric curves describing accuracy as a function of distance between target and comparison morph will be negatively correlated with participants' pro-white IAT scores.

h. (E) The influence of implicit race biases on performance metrics and signal detection parameters will be independent of the influence of explicit biases. 4. Black/White Face Recognition Memory Task (bwFRMT)

a. (C) Participant overall recognition memory accuracy will be significantly above chance (50%).

b. (C) Performance analyses ingroup/outgroup: Recognition memory accuracy will be higher for ingroup, as compared to outgroup, targets.

c. (E) Performance analyses implicit bias: Differences in recognition memory accuracy for white minus black targets will be positively correlated with participants'

pro-white implicit race bias.

d. (C) Signal Detection analyses ingroup/outgroup: Participants will show no difference in overall sensitivity for remembering outgroup versus ingroup faces, but will show decreased criterion thresholds for remembering outgroup compared to ingroup faces.

e. (E) Signal Detection analyses implicit bias: Participants' pro-white (i.e., white minus black) IAT scores will have a positive relationship with differences in

criterion thresholds for remembering white minus black faces. There will be no relationship between participants' IAT scores and overall sensitivity differences for remembering white minus black faces.

f. (E) The influence of implicit race biases on performance metrics and signal detection parameters will be independent of the influence of explicit biases.

3. Describe the key dependent variable(s) specifying how they will be measured. (optional)

The key dependent variables will be participant responses in the Race Working Memory Task (accuracy and reaction time; see task description below). The protocol will be administered online, implemented using Qualtrics (Qualtrics, Provo, UT) and jsPsych (de Leeuw, 2015), and hosted on pavlovia.org. Participants will be recruited via prolific.co and balanced with respect to gender and race. Upon arrival at the introductory web page, participants will complete a series of demographic questions and then proceed to the instructions for the Race Working Memory Task (RWMT). After completing the RWMT, participants will complete the Implicit Association Test, the Black/White Face Perceptual Discrimination Task, and the BLack/White Face Recognition Memory Task. Participants will then complete questionnaires assessing explicit race bias (race feeling thermometers; Axt, 2017), as well as the internal and external motivation to respond without prejudice scales (IMS/EMS; Plant & Devine, 1998)). Finally, participants will be debriefed as to the nature of implicit biases and then returned to prolific.co for payment. Participants will be paid for their time at a rate of $10.50 per estimated hour.

Tasks:

Race Working Memory Task (RWMT): In the RWMT participants are evaluated for the

influence of implicit race bias on the rate and accuracy with which they recall faces of varying races (white and black american faces). At the beginning of the task, participants are briefed in detail. Following the instructions, participants complete a set of practice trials, and then complete the task. On each trial, the participants are first shown a sample array of faces (set size = 2-4; 'Sample' period) each of which is paired with a neutral (with respect to valence and black/white stereotypicality) word describing a trait, and displayed for 2.5 to 5 seconds (depending on set size; 1.25 secs per face-trait pair). Trait words were selected from a list of 638 positive/neutral/negative personality traits (http://ideonomy.mit.edu/essays/traits.html) and the 6 most neutral and astereotypical (with regards to Black/White racial stereotypes) identified in a pilot study. Then the participant is shown a fixation cross for 4.0 seconds ('Delay' period). This is followed by a 'Match' period, in which the participant is presented with a single face-trait pair and asked to indicate whether the exact pair is a 'Match' (click 'e') to one of the pairs in the sample array or whether there is 'No Match' (click 'i'); participants have unlimited time to respond. Non-matches are novel face-trait pairings using the faces and traits from the 'Sample' period (i.e., neither the face nor the trait is novel). Participants will complete 96 trials (48 Black face trials and 48 White face trials). The independent variables are condition (Black/White faces, set size, match type) and the dependent variables are accuracy and reaction time (for correct responses only).

Implicit Association Test: The IAT is a measure of hidden biases, unconscious attitudes, and automatic preferences determined by measuring the time that takes an individual to classify concepts into two categories. Participants are shown 200 trials (including single-category and dual-category types). There are a total of 120 dual category trials of interest in which participants will be sorting words into Good and Bad categories, and faces into Black and White categories, simultaneously. Two blocks (20 and 40 trials respectively) will have a 'congruent' mapping (Good with White, Black with Bad), and two other blocks (20 and 40 trials respectively) an 'incongruent' mapping (Good with Black, White with Bad) ([in]congruency is defined with regard to stereotype norms in the U.S.). The resulting standard measure (calculated according to the algorithm in Lane et al., 2007) is called the "IAT D score" which ranges from -2 to 2, with positive scores indicating pro-white bias.

Black/White Face Perceptual Discrimination Task (bwFPDT; Hughes et al., 2019): The bwFPDT evaluates how sensitive participants are to variation in face similarity across different races. Stimuli consist of 8 Black and 8 White faces that were used to make 4 Black and 4 White face morph continua (in morph increments of 10%). Each trial begins with a fixation cross (duration=350ms). Then the participants are shown two faces (duration=500ms) from the same morph continua; the 50% morph (32 trials) and another morph that is either identical (32 trials) or different (ranging from 10% to 50% different; 80 trials). After 500ms, the faces will be replaced by a fixation cross, and participants will indicate via key press whether the 2 faces were identical or different. Participants will complete a total 112 trials. The independent variables are condition (black/white faces, degree of morph) and the dependent variables are accuracy and reaction time (for correct responses only).

Black/White Face Recognition Memory Task (bwFRMT; Hughes et al, 2019): The bwFRMT functions as a means of evaluating the participants' recognition memory and whether it varies between races. In the recognition memory task, 80 faces (40 White, 40 Black) will be used as stimuli. In the encoding phase, participants will be shown a sequence of 40 faces (20 White, 20

Black; 2secs per face) selected at random from the pool of 80. Participants will be instructed to memorize the faces, as they will answer questions about them later. Following the encoding phase, participants will complete a 3 minute distractor task in which they are asked to identify differences between 2 photographs. participants will then proceed to the test phase, in which they will be presented with each of the 80 faces in random order. Participants will indicate via key press whether each target is 'Old' (i.e., presented during the encoding phase) or 'New'. The independent variables are condition (black/white faces) and the dependent variables are accuracy and reaction time (for correct responses only).

Explicit race bias measures:

Internal/External Motivation to Respond Without Prejudice Scales (IMS-EMS): ;The IMS-EMS scales (Plant & Devine, 1998) consist of prompts focusing on the extent to which individuals are internally or externally motivated to not appear prejudiced towards 'Black

people'. The participants indicate the degree to which they agree with 10 statements on a nine point scale that ranges from 'Strongly Disagree' to 'Strongly Agree'.

Race feelings thermometers: The race feelings thermometer (see Axt, 2018) asks the participants to indicate on a scale from 0 ('Very Cold or Unfavorable') to 100 ('Very Warm and Favorable') their feeling towards Black/African Americans and White/Caucasian Americans.

Trait Stereotypicality and Neutrality Ratings(TSNR): Participants will also be asked to rate the Black/White stereotypicality and valence of 96 personality traits (32 each of Positive, Neutral, and Negative valence; taken from a list of 638 personality traits ≠ see http://ideonomy.mit.edu/essays/traits.html) including the 6 traits used in the rFW-WMT. Participants will rate the Black/White stereotypicality of each trait on a scale of 1 ('Strongly associated with Black Americans') to 7 ('Strongly associated with White Americans'), and the valence of the same word on a scale of 1 ('extremely negative') to 7 ('extremely positive').

4. How many and which conditions will participants be assigned to? (optional)

In this within-participant design, All participants will complete all of the tasks and questionnaires. Similarly, all participants will be exposed to all task conditions. The main tasks and conditions of interest are as follows: Race Working Memory Task, (white face arrays and the black face arrays, and array sizes); Black/White Face Perceptual Discrimination Task (morphs levels and race of morph); and Black/White Face Recognition Memory Task (race of face stimuli).

5. Specify exactly which analyses you will conduct to examine the main question/hypothesis. (optional)

Analyses

For all correlation analyses, we will calculate Pearson's r, the p-value (two-tailed), and the 95% bootstrapped confidence intervals. P-values will be compared to an alpha of 0.05 (adjusted for multiple comparisons when necessary) and bootstrapped confidence intervals

will be examined to determine if they exclude 0.

For all mean score and mean difference score analyses, A single sample t test (two-tailed) will be performed and 95% confidence intervals will be calculated. P-values will be compared to an alpha of 0.05 (adjusted for multiple comparisons when necessary) and bootstrapped confidence intervals will be examined to determine if they exclude 0.

1. Racial Face-Word Association Working Memory Task (rFW-WMT):

a. To address Hypothesis 4a- We will calculate mean recognition memory accuracy and test whether it is different from chance (i.e., 50%).

b. To address Hypothesis 4b - For each performance metric (i.e., reaction time and accuracy) we will perform a one-way ANOVA with face array size as the independent variable and performance as the dependent variable. If the overall F-statistic is significant, post-hoc comparisons will be used to determine pairwise condition-related differences in mean working memory performance. **c.** To address Hypothesis 4c - Mean difference in participant accuracy (collapsed across face array sizes) will be calculated for ingroup compared to outgroup trials.

d. To address Hypothesis 4d - We will perform a correlation to determine the relationship between IAT scores and the difference in response reaction times for white minus black faces (collapsed across face array sizes).

e. To address Hypothesis 4e - We will perform a correlation to determine the relationship between IAT scores and the difference in accuracy for white minus black faces (collapsed across face array sizes).

f. To address Hypothesis 4f - For each performance metric (i.e., reaction time and accuracy) we will perform a two-way ANOVA with face array size (1,2,3) and group membership (ingroup/outgroup) as the independent factors and performance as the dependent variable. Main effects and interactions will be tested. For any significant effects, post-hoc comparisons will be used to determine pairwise condition-related differences in working memory performance.

g. To address Hypothesis 4g - We will repeat the above analyses using signal detection measures (i.e. sensitivity and criterion threshold).

h. To address Hypothesis 4h - We will redo all the correlation analyses as a multiple regression with an explicit race bias and implicit race bias as predictors to account for the potential influence of implicit race bias.

2. IAT

a. To address Hypothesis 1a, we will calculate average IAT-D scores (with bootstrapped 95% confidence intervals) for each of the groups (White and Black American participants) and compare them to predictions.

3. Perceptual Discrimination Task

a. To address Hypothesis 2a - We will calculate the mean difference scores of ingroup (reaction time and accuracy) minus outgroup (reaction time and accuracy).

b. To address Hypothesis 2b - We will perform a correlation to determine the relationship between IAT scores and the difference in response reaction times for white minus black faces.

c. To address Hypotheses 2c - Signal detection theory values of sensitivity and criterion threshold will be calculated for each participant. We will then calculate the mean difference scores of ingroup (overall sensitivity and criterion threshold) minus outgroup (overall sensitivity and criterion threshold) faces and test that they are significantly different from 0.

d. To address Hypotheses 2d - Signal detection theory values of sensitivity and criterion threshold will be calculated for each target race condition

(Black/White) for each participant. We will then perform a correlation to determine the relationship between the differences in criterion threshold and sensitivity for White minus Black faces and participants IAT scores.

e. To address Hypothesis 2e - Participant accuracy data will be fit with a sigmoid as a function of morph distance. We will then calculate the mean of sigmoidal slopes across participants and test that it is significantly greater than 0.

f. To address Hypothesis 2f - Participant accuracy data will be fit separately for ingroup and outgroup trials with a sigmoid as a function of morph distance and the difference in slopes calculated. We will then calculate the mean of the difference in slopes across participants and test that it is significantly different from 0.

g. To address Hypothesis 2g - Participant accuracy data will be fit separately for the morph-race-related threshold differences with a psychometric curve as a function of distance between target and comparison morph will be calculated. We will then perform a correlation to determine the relationship between the morph-race-related threshold differences and participants' IAT scores.

h. To address Hypothesis 2h - We will redo all the correlation analyses as a multiple regression with an explicit race bias and implicit race bias as predictors to account for the potential influence of implicit race bias.

4. Race Recognition Memory Task

a. To address Hypothesis 3a - We will calculate mean recognition memory accuracy and test whether it is different from chance (i.e., 50%).

b. To address Hypothesis 3b - We will calculate the mean difference scores of ingroup (recognition memory accuracy) minus outgroup (recognition memory accuracy).

c. To address Hypothesis 3c - We will perform a correlation to determine the relationship between IAT scores and the difference in recognition memory accuracy for white minus black faces.

d. To address Hypothesis 3d - Signal detection theory values of sensitivity and criterion threshold will be calculated for each participant. We will then calculate the mean difference scores of ingroup (overall sensitivity and criterion threshold) minus outgroup (overall sensitivity and criterion threshold) faces.

e. To address Hypothesis 3e - Signal detection theory values of sensitivity and criterion threshold will be calculated for each target race condition

(Black/White) for each participant. We will then perform a correlation to determine the relationship between the differences in criterion threshold and sensitivity for White minus Black faces and participants IAT scores.

f. To address Hypothesis 3f - We will redo all the correlation analyses as a multiple regression with an explicit race bias and implicit race bias as predictors to account for the potential influence of implicit race bias.

6. Any secondary analyses? (optional)

Additional analyses will be performed to investigate whether the relationship between working memory and implicit race bias influences the signal detection threshold experienced in the Race Face-Word Working Memory Task. An additional analysis would be to investigate whether ingroup and outgroup biases, examined in the Perceptual Discrimination Task, influence the signal detection threshold experienced in the Race Face-Word Working Memory Task.

7. How many observations will be collected or what will determine the sample size?

Funding was provided for 150 participants. We performed a power analysis for each test type to determine whether 150 participants would generate a medium effect size at 80% power.

1. G Power Analyses

a. Single sample t-test - Means: difference from constant (one sample case)

i. At 80% power we can detect 0.5 effect size (cohen's d) with 34 participants.

ii. We have the sensitivity to detect effect sizes (cohen's d) of 0.230 at 80% power with 150 participants.

b. Correlation - point biserial

i. At 80% power we can detect 0.3 effect size with 82 participants.

ii. We have the sensitivity to detect effect sizes of 0.224 and above at 80% power with 150 participants.

c. Linear multiple regression

i. At 80% power we can detect 0.15 effect size (f^2) with 55 participants. ii. We have the sensitivity to detect effect sizes (f^2) of 0.053 at 80% with 150 participants.

d. One-way ANOVA - 1 group 3 measures

i. At 80% power we can detect 0.25 effect size (f) with 28 participants. ii. We have the sensitivity to detect effect sizes (f) of 0.104 at 80% with 150 participants.

e. Two-way ANOVA

i. At 80% power we can detect 0.25 effect size (f) with 19 participants. ii. We have the sensitivity to detect effect sizes (f) of 0.085 at 80% with 150 participants.

f. ANCOVA

i. At 80% power we can detect 0.25 effect size (f) with 269 participants. ii. We have the sensitivity to detect effect sizes (f) of 0.339 at 80% with 150 participants.

8. Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?) (optional)

Lack of Engagement: If participants give inaccurate responses for more than 50% of a task, reply to the questions with an unusually fast reaction time (<300ms) or only select a single response repeatedly for many trials (i.e. indicating 15 sequential face-words pairs are a match, their data will be discarded.(see Stanley et al, 2011).

In addition, the data from participants and trials that meet the following criteria will be excluded:

Participants who exhibit evidence of lack of engagement with the implicit measures of bias (IAT) will be excluded from the analyses featuring IAT measures. Lack of engagement will be established with unusually fast (<300ms) or slow (>10,000ms) reaction times as well as unusually repetitive response behavior or an unusually high error rate (IAT; see Lane et al, 2007 and Stanley et al, 2011).

In addition to examining the relationships between individual survey measures of explicit bias (Contact Measures, MRS, SRS, IMS/EMS) and our primary variables of interest, we will use component analysis (e.g. factor analysis or ICA) on survey measures of explicit attitudes (Contact

Measures, MRS, SRS, IMS/EMS) to reduce the inevitable redundancy between these measures, for use as covariates in other analyses.

References

638 Primary Personality Traits. (n.d.). Massachusetts Institute of Technology. Retrieved April 29, 2023, from http://ideonomy.mit.edu/essays/traits.html

Axt, J. R. (2018). The Best Way to Measure Explicit Racial Attitudes Is to Ask About Them. *Social Psychological and Personality Science*, *9*(8), 896–906. https://doi.org/10.1177/1948550617728995

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, *25*(9), 1804–1815. https://doi.org/10.1177/0956797614543801 Baars, B. J., & Gage, N. M. (2010). Chapter 2—A framework. In B. J. Baars & N. M. Gage (Eds.), *Cognition, Brain, and Consciousness (Second Edition)* (pp. 32–61). Academic Press. https://doi.org/10.1016/B978-0-12-375070-9.00002-4

Baddeley, A., Hitch, G., & Allen, R. (2020). C2A Multicomponent Model of Working Memory. In R. Logie, V. Camos, & N. Cowan (Eds.), *Working Memory: The state of the science* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*, 192–212. https://doi.org/10.1016/j.jesp.2005.07.003

Blough, D. S. (1959). Delayed matching in the pigeon. *Journal of the Experimental Analysis of Behavior*, *2*(2), 151–160. https://doi.org/10.1901/jeab.1959.2-151

Brady, T. F., Stormer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli | PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7459–7464. https://doi.org/10.1073/pnas.1520027113

Brown, T. I., Uncapher, M. R., Chow, T. E., Eberhardt, J. L., & Wagner, A. D. (2017). Cognitive control, attention, and the other race effect in memory. *PLOS ONE*, *12*(3), e0173579. https://doi.org/10.1371/journal.pone.0173579

Bowdler, J., & Harris, B. (2023, May 18). *Racial Inequality in the United States*. U.S. Department of the Treasury.

https://home.treasury.gov/news/featured-stories/racial-inequality-in-the-united-states Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57.

https://doi.org/10.1177/0963721409359277

Cowan, N., Blume, C. L., & Saults, S. (2013). *Attention to attributes and objects in working memory.* https://psycnet.apa.org/fulltext/2012-22642-001.html

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.

https://doi.org/10.1016/s1364-6613(03)00197-9

Darwin, C. J., & Baddeley, A. D. (1974). Acoustic memory and the perception of speech. *Cognitive Psychology*, *6*(1), 41–60. https://doi.org/10.1016/0010-0285(74)90003-6

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y Devine, P. G. (2001). Implicit prejudice and

stereotyping: How automatic are they? Introduction to the special section. *Journal of Personality and Social Psychology*, *81*(5), 757–759. Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*, 978–980.

https://doi.org/10.1111/j.1467-9280.2008.02186.x

Estes, W. K. (2014). *Handbook of Learning and Cognitive Processes (Volume 6): Linguistic Functions in Cognitive Theory*. Psychology Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Ferster, C. B. (1960). Intermittent reinforcement of matching to sample in the pigeon. *Journal of the Experimental Analysis of Behavior*, *3*(3), 259–272. https://doi.org/10.1901/jeab.1960.3-259 Fiez, J. A. (2016). Chapter 68—Neural Basis of Phonological Short-Term Memory. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 855–862). Academic Press.

https://doi.org/10.1016/B978-0-12-407794-2.00068-7

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998a). 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., McGhee, D. E., & Schwartz, J. L. K. (1998b). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Gross, S. R., Possley, M., Otterbourg, K., Stephens, K., Paredes, J., & O'Brien, B. (2022). Race and Wrongful Convictions in the United States 2022. *SSRN Electronic Journal*.

https://doi.org/10.2139/ssrn.4245863

Hawk, J., & Abel, T. (2010). Role of Gene Transcription in Long-Term Memory Storage. In G. F. Koob, M. L. Moal, & R. F. Thompson (Eds.), *Encyclopedia of Behavioral Neuroscience* (pp. 161–179). Academic Press. https://doi.org/10.1016/B978-0-08-045396-5.00030-0

Henry, L. A. (2010). The episodic buffer in children with intellectual disabilities: An exploratory study. *Research in Developmental Disabilities*, *31*(6–3), 1609–1614. https://doi.org/10.1016/j.ridd.2010.04.025

Hughes, B. L., Camp, N. P., Gomez, J., Natu, V. S., Grill-Spector, K., & Eberhardt, J. L. (2019). Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(29), 14532–14537. https://doi.org/10.1073/pnas.182208411

Hupbach, A., Gomez, R., & Nadel, L. (2009). Episodic memory reconsolidation: Updating or source confusion? *Memory (Hove, England)*, *17*(5), 502–510. https://doi.org/10.1080/09658210902882399

Implicit race attitudes predict trustworthiness judgments and economic trust decisions | *PNAS*. (n.d.). Retrieved April 29, 2023, from https://www.pnas.org/doi/abs/10.1073/pnas.1014345108 Keisler-Starkey, K., & Bunch, L. N. (n.d.). *Health Insurance Coverage in the United States: 2021*. Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The Price of Racial Bias: Intergroup Negotiations in the Ultimatum Game. *Psychological Science*, *24*(12), 2498–2504. https://doi.org/10.1177/0956797613496435

Lane, K., Banaji, M. R., & Nosek, B. A. (2007). Implicit Measures of Attitudes. *Guilford Press*, 59–102.

Logie, R., Camos, V., & Cowan, N. (2020). *Working Memory: The state of the science*. Oxford University Press.

Match-to-sample task. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Match-to-sample_task&oldid=1091167720

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The Relationship Between Working Memory Capacity and Executive Functioning: Evidence for a Common Executive Attention Construct. *Neuropsychology*, *24*(2), 222–243. https://doi.org/10.1037/a0017619

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual

Working Memory in Prefrontal Cortex of the Macaque. *Journal of Neuroscience*, *16*(16), 5154–5167. https://doi.org/10.1523/JNEUROSCI.16-16-05154.1996

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*(8), 831–864. https://doi.org/10.1037/bul0000051

Morris, R., Hitch, G., Graham, K., & Bussey, T. (2006). CHAPTER 9—Learning and Memory. In R. Morris, L. Tarassenko, & M. Kenward (Eds.), *Cognitive Systems—Information Processing Meets Brain Science* (pp. 193–235). Academic Press.

https://doi.org/10.1016/B978-012088566-4/50015-5

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738.

https://doi.org/10.1162/089892900562552

Phillips, I. B. (2011). Perception and Iconic Memory: What Sperling Doesn't Show. *Mind & Language*, *26*(4), 381–411. https://doi.org/10.1111/j.1468-0017.2011.01422.x

Pollmann, S., & Schneider, W. X. (2022). Chapter 20 - Working memory and active sampling of the environment: Medial temporal contributions. In G. Miceli, P. Bartolomeo, & V. Navarro (Eds.), *Handbook of Clinical Neurology* (Vol. 187, pp. 339–357). Elsevier.

https://doi.org/10.1016/B978-0-12-823493-8.00029-8

Reihl, K. M., Hurley, R. A., & Taber, K. H. (2015). Neurobiology of Implicit and Explicit Bias: Implications for Clinicians. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *27*(4), A6-253. https://doi.org/10.1176/appi.neuropsych.15080212

Riley, W. J. (2012). Health Disparities: Gaps in Access, Quality and Affordability of Medical Care. *Transactions of the American Clinical and Climatological Association*, *123*,

167–174. Roy, R. N., Bonnet, S., Charbonnier, S., & Campagne, A. (2013). Mental fatigue and working memory load estimation: Interaction and implications for EEG-based passive BCI. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6607–6610. https://doi.org/10.1109/EMBC.2013.6611070

Schwartz, G. L., & Jahn, J. L. (2020). Mapping fatal police violence across U.S. metropolitan areas: Overall rates and racial/ethnic inequities, 2013-2017. *PLOS ONE*, *15*(6), e0229686. https://doi.org/10.1371/journal.pone.0229686

Skinner, B. F. (19510401). Are theories of learning necessary? *Psychological Review*, *57*(4), 193. https://doi.org/10.1037/h0054367

Vallar, G. (2017). Short-Term Memory☆. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.03170-9

Wang, M., Gamo, N. J., Yang, Y., Jin, L. E., Wang, X.-J., Laubach, M., Mazer, J. A., Lee, D., & Arnsten, A. F. T. (2011). Neuronal basis of age-related working memory decline. *Nature*, *476*(7359), 210–213. https://doi.org/10.1038/nature10243

Westbrook, J. I., Raban, M. Z., Walter, S. R., & Douglas, H. (2018). Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: A prospective, direct observation study. *BMJ Quality & Safety*, *27*(8), 655–663. https://doi.org/10.1136/bmjqs-2017-007333

Addendum A