# FORGOTTEN BUT NOT FORGIVEN: THE RETENTION OF IMPLICIT NEGATIVE BIAS DESPITE EXPLICIT FORGETTING

A Thesis

Presented to the Faculty of the Graduate School

Of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts

by

Christine Bell Johnson

August 2021

# ABSTRACT

Previous studies have shown that implicit attitudes are much harder to reverse with updated information than are explicit attitudes. Yet, very little research has been conducted on a phenomenon we are calling “forgotten but not forgiven.” That is, if we accuse someone of wrongdoing, then subsequently exonerate them by discrediting the information presented, will people who have been exposed to negative information about the character hold onto a negative implicit bias against the target even when they fail to recognize the person as the target on a conscious, explicit level? Our participant pool was comprised of 170 Cornell University students, who were shown information, told it was untrue, and took an Implicit Association Test. The main analysis was a t-test of participants’ D-score, which was not statistically significant. Remembrance rates of the target remained low, however, and we suggest continued research on this topic because of certain limitations we faced with the COVID-19 pandemic.

Christine Bell Johnson

# BIOGRAPHICAL SKETCH

Christine Johnson graduated from Cornell University in May of 2020 with a Bachelor's of Science in Development Sociology with minors in Law and Society, Inequality Studies, and Demography. While pursuing her undergraduate degree, she conducted research with the National Science Foundation on topics of impulse-control disorders, general mental health, and self-rated health in Ukraine populations. Christine began work with the Child Witness and Cognition Lab her concluding semester as an undergraduate, conducting research on persistent implicit biases, on which the current paper expands. Joining the Cornell Department of Human Development in August of 2020 to pursue a Master's in Developmental Psychology, Christine holds a concentration in Law, Psychology, and Human Development and her research focuses on implicit biases that persist even after a person has been exonerated and, indeed, forgotten.

# DEDICATION

For my parents, Julie and Karl, who always support and encourage me: thank you for believing in me and helping me reach my goals.

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my faculty advisor, Dr. Stephen J. Ceci. Completing a graduate degree in the midst of a global pandemic brought its fair share of challenges, but Dr. Ceci was always ready to offer his expertise or a word of heartfelt encouragement when I needed it the most. It is because of Steve's abundant support that this thesis was made possible. I would also like to thank Dr. Wendy Williams for agreeing to be my secondary committee member and for her supporting role in the completion of this degree. My gratitude also to Andrew Baron of the University of British Columbia, for his contribution in permitting our use of CloudArmy's IAT software.

A special thank you to Cornell/Yale Ph.D. student Reut Shachnai for lending her experience to every step of this research. I would also like to express my appreciation to my undergraduate research assistant, Alexis Aberman, for her invaluable contributions.

And to Christopher, my fellow graduate student and my best friend, thank you for bringing so much joy and light into my life and inspiring me in everything I do.

# TABLE OF CONTENTS

Biographical Sketch..........iii

Dedication..........iv

Acknowledgments..........v

Table of Contents..........vi

Introduction..........1

Method..........9

Results..........13

Discussion..........18

References..........20

# INTRODUCTION

"You can only make a first impression once." While this is, in principle, a factual statement (and a popular saying when meeting one's potential future in-laws), it is also true that actively revising previously formed attitudes is a common adaptation of everyday life. Many of us have close friends that we disliked at first meeting but grew to love. And yet others of us have enemies whom we once adored. Regardless, evaluation is one of the main methods by which we assess the world around us (Mann et al., 2015; Wilson et al., 2000). First impressions are a common form of evaluation, as they occur with all novel introductions, whether it be meeting someone entirely new or watching a movie for the first time. As humans, we constantly react to the world around us in a spontaneous and rapid fashion. A facet of this is our seemingly robust instinct to make snap judgements of others based on extremely limited information, an action we cannot easily inhibit (Gladwell, 2005; Todorov et al., 2013; Todorov, 2017; Winter et al., 1985). Moreover, this conscious and explicit liking or disliking towards the attitude object is accompanied by an implicit and routine positive or negative reaction (Gregg et al., 2006). It is important to distinguish precisely what we mean by explicit and implicit attitudes.

If you were to be asked what you thought of someone, your response would reflect your explicit attitude. We are conscious of, and have the ability to control the way in which we express, our explicit attitudes (Fazio, 1995; Rydell et al., 2006). Your implicit attitude of a person is certainly harder to determine. In fact, much of the time people have little to no awareness of their true implicit attitudes. That is to say, implicit evaluations are *unintentional* and often, but not always, unconscious (Blair et al., 2015). Thus, researchers have come to use the Implicit Association Test (IAT) in an attempt to reveal the implicit, often hidden, attitudes of the test-taker (Greenwald et al., 1998).

## The Implications of First Impressions

Whether our first impressions are good or bad, they certainly have wide-reaching implications. This is in part due to our inclination to pay the greatest attention to evidence that confirms our presently held beliefs, whilst readily dismissing information that fails to espouse these convictions; this has been widely researched in regards to political attitudes, which are often marked by misinformation (Bartels, 2002; Munro, 1995; Thorson, 2015). This has especially treacherous consequences in the legal system, where false confessions have been shown to lead to a significant number of subsequent evidence errors (Kassin et al., 2012). In many facets of life, we fall victim to the confirmation bias and are inclined to falsely interpret new information as supporting our previously held beliefs, despite evidence that contradicts them (Dougherty et al., 1994; Lewandowsky et al., 2012; Nyhan & Reifler, 2010; Porter et al., 2010; Rabin & Shrag, 1999).

If initial judgements are robust and consequential as such, what happens when first impressions are based on false information? Certainly, we encounter numerous inconsistencies between hearsay and reality, no matter how small, on a regular basis. After all, we live in an age of "information overload;" much of this information, however, is disinformation or misinformation, sometimes referred to as "fake news" (Levitin, 2014). Accusations of sexual misconduct inarguably plague a stunning number of politicians and other public figures of our time. Is redemption possible after a person's reputation takes a turn for the worse? This question is of great importance, as misinformation saturates our everyday lives and often causes destruction to careers and livelihoods too little too late. If false information is so rampant, should this not require us to update and revise our opinions and beliefs about certain subjects and people as we learn and unlearn—or discredit prior—information? Revision of former understanding is certainly necessary in principle, though the minutiae of how, and how well, we are able to accomplish this are of the utmost importance. Allegations based on disproven accounts are undoubtedly harmful even *after* a person is exonerated. This is because when attitudes are changed due to updated knowledge, the original impression of the person is typically not completely erased, allowing memories of the past to impact ensuing behavior towards, and attitudes about, that person (Gawronski & Strack, 2004; Gregg et al., 2006; Petty et al., 2006; Wilson et al., 2000). Research strongly suggests that this phenomenon, often called the Continued Influence Effect, belief echoes, or the PAST (Past Attitudes Are Still There) model, holds steady in many circumstances of information revision (Johnson & Seifert, 1994; Petty et al., 2006; Thorson 2015).

## Dual-Processing Models

Importantly, not only do past impressions have a continued influence through implicit means, but we also consider the possibility that the systems that form and maintain implicit evaluations may function autonomously from a person's conscious, self-recognized sentiments. That is, certain evidence suggests that explicit and implicit processes are separate from one other, this duplicity allowing a person to have an implicit attitude that directly contradicts their explicit attitude (Cone & Ferguson, 2015; Rydell et al., 2006; Sloman 1996; Smith & Decoster, 2000; Wilson et al., 2000). Sloman (1966) argued that we have two distinct learning processes, one for implicit attitudes and the other for explicit. Wilson et al. (2000) found this model of dual attitudes to be reflected in many aspects of cognition such as memory, motivation, and self-esteem. However, dual attitudes were found to have the strongest relation to prejudice, or bias, providing further support for the ability to have two conflicting attitudes, and thus a lingering negative bias, towards a single object. If multiple attitudes, old and new, explicit and implicit, can exist simultaneously within memory, then unconscious attitudes constantly impact a person’s responses to stimuli, even despite a person's cognizance of the opposite attitude. Implicit attitudes are activated automatically whereas explicit attitudes require more effort to retrieve, the attitude a person affirms at a given time being a function of the cognitive capacity they have available at the time, and whether the explicit attitude dominates in salience or is less accessible (Wilson et al., 2000).

It is important to note that though dual processes theory has been replicated, it has also gained conflicting evidence. Peters and Gawronski (2011) also hypothesized that explicit and implicit attitudes are a direct result of dual learning processes, however they found evidence in contradiction to this hypothesis. Others have argued that there is not enough evidence to accept the dual-processing model and that experiments that have posited this have yet to be properly evaluated (Mitchell et al., 2009).

## Attitude Revision

Whether or not a dual-processing model exists, it is clear that when information given about a person or group leads to a negative evaluation, *explicit* attitude revision is possible when the information is subsequently retracted or reframed to remove or discredit the negative impression. Mann & Ferguson (2015) provide a good example of this in the laboratory when they relayed a story to participants about a man who broke into homes and brought about considerable damage in doing so. Then, the information is reframed to explain his actions as an attempt to save children from house fires. This allows for reinterpretation of the prior knowledge and thus a reversal of a person’s formerly negative explicit attitude towards the character in the story (Mann & Ferguson, 2015). Accordingly, we can say with relative certainty that conscious attitudes can quite often be reversed as is the case when the villain of a story becomes the hero or vice versa. In fact, an experiment by Rydell & McConnell (2006) exemplifies that exposure to a minor amount of counter-attitudinal information can effectively alter explicit attitudes. Despite the seemingly malleable characteristic of explicit attitudes, empirical evidence strongly suggests the reversal of implicit attitudes is much harder to accomplish (Mann & Ferguson, 2015; Petty et al., 2006; Rydell & McConnell, 2006). This is especially true when there is little or no context for the information reversal (besides simply labeling the information as untrue) or when a person is experiencing any level of cognitive load when processing updated information (Mann & Ferguson, 2015; Wilson et al., 2000). Generally, automatic and implicit attitudes are more resistant to change subsequent to their formation than are self-reported, explicit attitudes. That is to say, automatic evaluations are asymmetrically malleable: more easily formed than they are reversed, Gregg et al. (2006) likens this to gaining weight or accruing debt (Mann & Ferguson, 2015; Petty et al., 2006; Rydell & McConnell, 2006).

Implicit evaluations transform more efficiently and thoroughly in the negative direction: accusing a widely respected person of sexual misconduct or another misdeed can immediately taint implicit attitudes towards the accused, attitudes that were once undoubtedly positive. This is due to the negativity bias, which posits that negative information is treated as significantly more diagnostic (and thus more reliably representative of a person’s character) compared to positive information (Cone & Ferguson, 2015; Ito et al., 1998). Considering this, it is thought that changing a person’s opinion of someone else from negative to positive is quite harder than the opposite: to most people, one good quality is not redeeming, but one bad quality can certainly be damning (Baumeister et al., 2001; Fiske, 1980; Levey & Martin, 1975; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). Because of this, it is important to be extra vigilant when making accusations and convictions. This is especially imperative in a time where news, true and otherwise, spreads quickly and irreparably.

## Forgetting but not Forgiving

Our aim is to discover whether an implicit bias or subconscious negative attitude can persist even past a person's conscious memory. Victorino et al. (2019) conducted a similar experiment in two parts, the first of which asked participants to rate a piece of chocolate and the second, a scientific paper, all the while priming them with the name of a university that they claimed was funding their research study. The priming stimuli across both experiments were footnotes indicating that funding for the research was of either European or African origin, depending on the condition assigned to the participant. It was determined that the priming stimuli, while not remembered by 92.5% of participants, still had a significant impact on the judgements and evaluations made by participants in both experiments. Specifically, participants showed significant bias against the African country prime in comparison to the European prime, as predicted (Victorino et al., 2019).

Bastick (2021) is also relevant to the present research because they demonstrated the ability for disinformation to change participants' behaviors without their conscious awareness. Participants took a Finger Tapping Test while being exposed to fake news, and though they were unaware of the changes in their behavior, participants' finger tapping increased significantly after exposure to positively-valenced fake news. This finding supports the current research's hypothesis that implicit biases can persist in participants even when they explicitly forget the target.

## The Present Study

The present study is similar to Victorino et al. (2019) and Bastick (2021) in that all three studies aimed to test the persistence of bias effects in the midst of forgetting. However, our work also differs from these studies in several key ways. First, it utilizes an IAT, which is designed to measure the strength of implicit associations between objects and evaluations. The IAT is thought to be a reliable assessment of implicit bias effects (Greenwald et al., 1998). Next, the present study facilitates forgetting by presenting many verbal descriptions and images briefly; rather than using a subtle prime, there are many primes and objects of evaluation that are shown in a short period of time. We also employ fictional people rather than objects, as biases against exonerated persons have often dire consequences. Finally, the present study is unique in that it investigates continued bias effects through an IAT after exposure to misinformation, which has important implications during a time in which misinformation is amplified by the media.

This study not only aims to replicate past research evaluating implicit attitudes, and particularly their resistance to change despite explicit attitude reversal, but we also seek to examine the role forgetting plays in reversing or failing to reverse negative implicit attitudes. According to past research, implicit attitudes are sticky and often persist for a long while after information has invalidated previous accusations. Importantly, implicit attitudes and biases are also likely to persist a long while after explicit, conscious attitudes have been corrected and reversed. In the current study, we aim to see if this implicit attitude persistence can exist even when the attitude holder is unable to recognize the person they are evaluating. That is, can they forget the person, and thus the accusations against the person, and yet still not forgive them on an implicit level? We refer to this phenomenon as “forgetting but not forgiving.” To test this, we present participants with a fictional person, i.e., “Person X,” and accuse them of wrongdoing, then subsequently exonerate Person X by invalidating the information presented. Notably, we do not reframe the information as most real-life false accusations are not awarded this nicety. We attempt to determine whether an individual who has been exposed to these statements about Person X is likely to hold on to a negative, unconscious (implicit) bias against Person X even while failing to identify them as the target on a conscious, explicit level. We suspect that a persistent negative attitude can exist on an implicit level even when a person does not recognize the target's face. Although we have no *a priori* predictions regarding outcomes of the following, we also explore whether gender or race correlate with implicit bias against the target.

# METHOD

## Participants and Materials

Participants were undergraduate students taking courses that offered credit for research participation at Cornell University. Before beginning to collect data, we conducted a G-power analysis on a small-to-moderate effect size of d=.2 at the usual power (.80) and alpha (0.05). A simple t-test between a bias effect and zero requires a sample of *N*=156. We recruited a sample of *N*=170 in anticipation that there would be participants who met the exclusionary criteria. We chose to use exclusionary criteria for any participants who (1) failed to complete the IAT in full, and/or (2) met one or more of the following IAT exclusion criteria set by Greenwald et al., (2003): (a) ⩾10% of responses faster than 300 ms, (b) error rate of ⩾35%, or (c) average response latency of 3 SD above the sample's mean response latency. No participant met the exclusionary criteria and thus all 170 were included in our analyses.

The 170 participants included 28 males, 140 females, and 2 persons who identified as another gender or preferred not to say. Of the 170 participants, there were 78 who described themselves as White, 62 as Asian, 12 as Black or African American, 9 as Hispanic or Latinx, and 9 who chose "Other" or preferred not to disclose their racial identity.

## Procedure

The study began by introducing participants to 12 different men. These men's photos were selected from the Chicago Face Database (Ma et al., 2015). In an attempt to reduce the rate of remembrance of the target by introducing too many distinguishing factors, we comprised the pool of faces to include only young looking, white males with varying hair colors, eye colors, and facial shapes and features. Each of these men are denoted to participants with various letters (i.e., “Person M”). We suspected the use of specific names could provide participants with associations if they knew a person or persons with the same name. Additionally, names can provide an extraneous reason for positive or negative judgement, depending on if the participant liked or disliked each name. The participants were randomly assigned to one of three different conditions at an equal rate, with the target faces alternating with each condition in anticipation of the potential for judgements of attractiveness and to control for these and other differentiating visual evaluations.

The participant started by reading a brief conversation between two college women who discuss a man, painting him in a positive, negative, or neutral light. When the participant finished reading the conversation, they continued to the next slide to see a photo of the corresponding man’s face with a 1-second exposure time. Twelve of these conversations and photos were shown to participants consecutively. We arrived at the number 12 after piloting the study multiple times and determining that exposure to fewer faces increased participants’ recognition of the target, as did a longer (multiple-second) exposure time.

Participants were shown several faces that acted as distractors before the target appeared, allowing them to understand the flow of the study and the brevity of each face’s appearance. The target appeared in the middle of the lineup as opposed to the end to reduce the chance for the recency effect to increase recognition of the target (Baddeley & Hitch, 1993). The neutral, or “control” face, which appeared in the IAT along with the target, was introduced to participants as one of the 12 faces and was described in a neutral way. By introducing the control’s face before the IAT, we avoid the potential for the mere exposure effect, a phenomenon whereby repeated exposure to a specific stimulus leads an individual to respond to that stimulus more positively, therefore having the potential to cause participants to be partial towards the target (Zajonc, 1968). Notably, the effect remains consistent when images of people are used as stimuli, as demonstrated by the strong relationship between familiarity and the perceived likability of the images of people (Harrison, 1969). Therefore, the risk that participants may experience a more positive attitude toward a specific, familiar face is mitigated by our consistent use of non-novel faces in the IAT.

Each of the 12 photos we introduce in the beginning of the study has a corresponding photo of a different face's photo that is its normed counterpart. Norming was conducted by the Chicago Face Database researchers and determined by both objective measures of facial features and subjective ratings completed by a pool of participants (Ma et al., 2015). We selected pairs by choosing photos from the database that were pre-determined by norming to be very similar to each other. The 24 total photos, consisting of the 12 faces introduced to the participants and each of their normed matches, appear in the recognition task at the end of the study. This is intended to determine whether the participant undoubtedly recognizes and remembers the target.

Gregg et al. (2006) discovered that participants in their study created instant implicit attitudes towards novel persons described as having either positive or negative attributes. We attempt something similar by exposing the participants to 12 different people described in a bad, good, or neutral way.

The target was depicted in a negative way and the control (who provided a comparison in the IAT) was described in a neutral fashion. Specifically, the target was accused of wrongdoing (Person T cheated on their last partner) along with a few other non-target persons to detract attention from the target. The participants then completed an abridged version of the Big 5 personality test, the Mini-IPIP, by rating how much they agree to different statements about their personality on a 7-point Likert scale. This acted not only as a distinction from previous studies on persistence of implicit attitudes but also as a distractor from the previously shown information to aid in forgetting the target explicitly. We suspected that once the participants began to think about themselves, they would begin to forget the details of what they had just read. The target was then exonerated by revealing to the participants that the previous information discussed regarding the 12 different people was all false information and therefore they should forget what was stated.

Participants were then given an Implicit Association Test (IAT) to determine their implicit attitudes towards the target in comparison to the neutral persons from the beginning of the study. The participants were assigned to one of six different IATs. For each of the three versions of the target exposure section of the study, two different Implicit Association Tasks were created, one which began with Person A on the right-hand side and Person B on the left-hand side, and the other which began with Person B on the right-hand side and Person A on the left-hand side at the start of the test. As with any conventional IAT the test goes through two cycles, the second of which is identical to the first aside from swapping the right and left-hand side photos for a reversal effect. Upon completion of the IAT, participants were asked to complete a memory test where 24 photos were shown. The participants were asked to select 12 of the 24 photos they remember best as being introduced to in the beginning of the survey. Of those 12, they then selected the one they believed to be the target photo and rated how confident they were that they chose the correct photo on a 5-point scale from “Not at all confident” to “Extremely confident.”

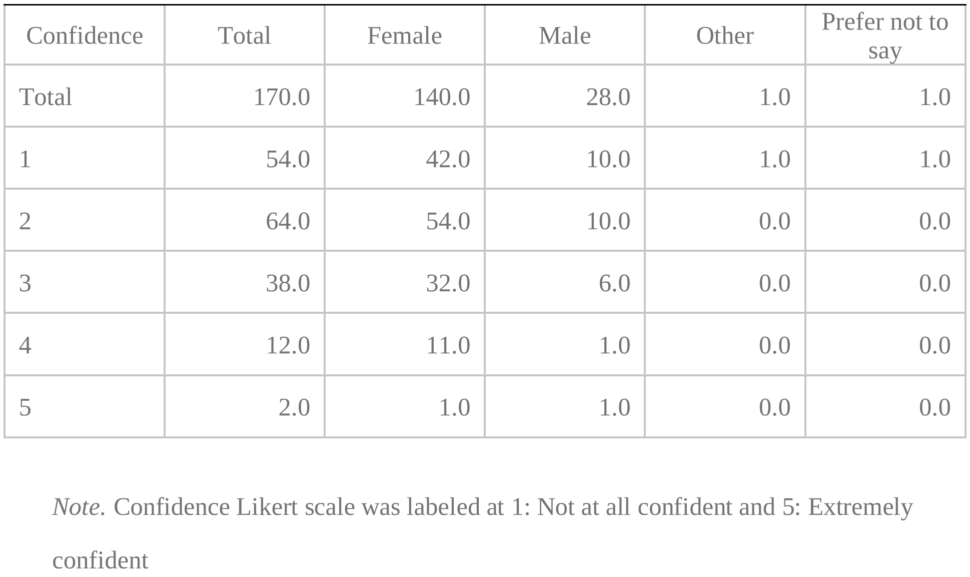
# RESULTS

## Target Non-Remembrance

The majority of participants were unable to correctly identify the target, as we had hoped. Only 24 out of the 170 participants correctly selected the target for their version of the study, meaning approximately 86% of participants failed to remember the target when given the memory test. While this is the outcome we hoped for, it is possible that the correct selections were largely a function of the faces' similarities as participant confidence levels for selecting the correct person were consistently low (see Table 1).

**Table 1**

**Participant Self-Report of Confidence for Selecting Target by Gender**



| Confidence | Total | Female | Male | Other | Prefer not to say |
|---|---|---|---|---|---|
| Total | 170.0 | 140.0 | 28.0 | 1.0 | 1.0 |
| 1 | 54.0 | 42.0 | 10.0 | 1.0 | 1.0 |
| 2 | 64.0 | 54.0 | 10.0 | 0.0 | 0.0 |
| 3 | 38.0 | 32.0 | 6.0 | 0.0 | 0.0 |
| 4 | 12.0 | 11.0 | 1.0 | 0.0 | 0.0 |
| 5 | 2.0 | 1.0 | 1.0 | 0.0 | 0.0 |

*Note.* Confidence Likert scale was labeled at 1: Not at all confident and 5: Extremely confident

The average confidence of the overall sample was a score of 2.1, and the median of the sample was a score of 2.0. These average and median scores for confidence were consistent across males and females, with the exception of the male average score being slightly lower at

2.0. Overall, the majority of respondents chose option (1) or (2), signaling a very low confidence in having chosen the correct target across most participants.

**Table 2**

**Participant Self-Report of Confidence for Selecting Target by Race**

| Confidence | Total | Black or African America n | White | Asian | Hispanic /Latinx | Other | Prefer not to say |
|---|---|---|---|---|---|---|---|
| 1 | 54.0 | 3.0 | 20.0 | 24.0 | 4.0 | 2.0 | 1.0 |
| 2 | 64.0 | 6.0 | 32.0 | 20.0 | 3.0 | 3.0 | 0.0 |
| 3 | 38.0 | 2.0 | 17.0 | 16.0 | 0.0 | 3.0 | 0.0 |
| 4 | 12.0 | 1.0 | 8.0 | 2.0 | 1.0 | 0.0 | 0.0 |
| 5 | 2.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| average | 2.1 | 2.1 | 2.2 | 1.9 | 2.1 | 2.1 | 1.0 |
| median | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 1.0 |

*Note.* Confidence Likert scale was labeled at 1: Not at all confident and 5: Extremely confident

Although the average and median scores of confidence were fairly consistent across race, those who identified as White reported being slightly more confident, with an average rating of 2.2 as compared to the total average of 2.1. Those who identified themselves as Asian were slightly less confident of having correctly identified the target than the overall average of 2.1, with an average score of 1.9. The samples of participants who self-identified as Black or African American, Hispanic or Latinx, or “Other” had median and average scores consistent with the median and average scores of the overall sample.

## Calculating D-Score

After determining that no participant met any of the exclusionary criteria, the D-score was calculated with the full sample of N=170. This was done by finding the difference between the Implicit Association Test reaction times towards the control and the reaction times towards the target, and dividing this difference by the standard deviation of within-subject pooled response times. We then conducted a one-sample t-test for the D-score (see Table 3).

**Table 3**

**One-Sample T-test for D-score**

| | |
|---|---|
| t | -1.469 |
| df | 169 |
| p-value | 0.1437 |
| 95% confidence interval | -0.09951159, 0.01459945 |

As shown in the table above, the t-test was non-significant with the value of t being –1.469 and a p-value of 0.1437. The negative value of t indicates a slight, non-significant trend in favor of the target as opposed to the control. As such, we did not find evidence to support our hypothesis that an exonerated person can still receive negative bias despite being forgotten.

In addition to our main analysis, we conducted exploratory analyses to examine the relationship between specific demographics and D-score. The first of these exploratory analyses was an analysis of variance (ANOVA) for gender, which we conducted as a means of examining whether there was a statistically significant difference in bias towards the target between genders of participants.

**Table 4**

**ANOVA for Gender by D-score**

| | Df | Sum Sq | Mean Sq | F value | P value |
|---|---|---|---|---|---|
| Gender | 3 | 21747 | 7249 | 0.997 | 0.396 |

As seen in Table 4, the analysis of variance of D-score for gender was non-significant with an F-value of 0.997. Thus, there is no sufficient evidence to suggest that bias towards the target differed between participants' genders.  It is important to note our relatively small sample size of N=28 males compared to N=140 females. The disproportionate male and female sample sizes make it difficult to argue for the external validity of this finding, though we hope subsequent research on this topic will consider gender differences as well.

It is possible that the large difference in male and female populations is the sole explanation for this finding, though it is also reasonable to conjecture that it is a fairly accurate reflection of implicit attitudes as being relatively consistent across gender.

**Table 5**

**ANOVA for Race by D-score**

| | Df | Sum Sq | Mean Sq | F value | P value |
|---|---|---|---|---|---|
| Race | 5 | 30605 | 6121 | 0.838 | 0.525 |

The analysis of variance of participants' D-scores for race was also non-significant with an F-value of 0.838. Across the 5 different races, no statistically significant difference in bias towards the target was found. We had no specific hypothesis regarding race at the start of this study, but rather sought to examine how differing demographics might play into this phenomenon of forgetting but not forgiving. It is conceivable to attribute the non-significance of race and gender to factors that we suspect also contributed to the non-significance of the D-score for the participant population as a whole (see Discussion), though it is also possible that attitudes of implicit bias do not significantly differ across these demographics.

# DISCUSSION

In the current study, we sought to test whether an exonerated person could be forgotten but still remain the subject of implicit bias. This analysis of bias turned out to be non-significant. The analyses of variance for both gender and race on bias were likewise both non-significant. Overall accurate recall of the target remained low with 24 of 170 participants correctly identifying the target. Confidence ratings for correctly identifying the target were also low, with an overall sample average of 2.1 out of 5.

There are a number of reasons that could potentially explain why we have arrived at this outcome. The first is that the brief exposure to the target was not lengthy enough to produce a salient dislike of the target, leading participants to have no real associations attached to the target's face while taking the IAT. Our biggest hurdle in the design of this study was in crafting a balance between exposure of the target that was long enough for participants to digest and form a negative impression, but not long enough for them to remember it explicitly throughout the duration of the study. It is also possible that the targets and controls were too similar in appearance, and when presented with both faces, the participant could not attribute an implicit positive or negative feeling to one rather than the other, or perhaps even confused the control as being the target. Although we were able to administer a few rounds of pilot testing, we had neither the time nor flexibility of in-person data collection to determine the best course of action through usability testing.

Finally, the restriction of data collection during COVID-19 (i.e., online) required that data would be collected in a single session. From memory research it might behoove us to separate the exposure of faces from the IAT, administering the latter several days late to give enough time for explicit memory to fade.

As stated previously, due to the COVID-19 pandemic, we were unable to collect data in the lab. We propose that the "forgotten but not forgiven" phenomenon is important enough to deserve subsequent study, as even corrected misinformation can result in the undeserved end to political careers, continued bias against the wrongly convicted, and much more. It is reasonable to think that a two-part study where participants are shown the target for more than 1 second, or multiple times, then given a few days or weeks to forget what the target looks like, and asked to come back to the lab to take the IAT would present a better opportunity to explore this phenomenon. Since it is plausible that the faces resembled each other so strongly that participants could not distinguish between them enough to subconsciously attribute negative feelings towards one rather than the other, it may be wise to use a more diverse set of persons to act as targets and controls in the study. We believe there is evidence to suggest that this phenomenon exists and can be tested accurately with the right approach as exemplified by both Victorino et al. (2019) and Bastick (2021).

# REFERENCES

Baddeley, A. D., & Hitch, G. (1993). The recency effect: implicit learning with explicit retrieval?: short-term memory. Memory & Cognition, 21(2), 146–155.

Blair, I. V., Dasgupta, N., & Glaser, J. (2015). Implicit attitudes. APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition., 665-691. doi:10.1037/14341-021

Bartels, L.M. Beyond the Running Tally: Partisan Bias in Political Perceptions. Political Behavior 24, 117–150 (2002). https://doi.org/10.1023/A:1021226224601

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. Computers in Human Behavior, 116. https://doi-org.proxy.library.cornell.edu/10.1016/j.chb.2020.106633

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. Review of General Psychology, 5(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. Journal of Personality and Social Psychology, 108(1), 37–57. https://doi.org/10.1037/pspa0000014

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales : Tiny-yet-effective measures of the big five factors of personality. Psychological Assessment, 18(2), 192–203.

Dougherty, T. W., Turban, D. B., & Callender, J. C. (1994). Confirming first impressions in the employment interview: a field study of interviewer behavior. Journal of Applied Psychology, 79(5), 659–665.

Fazio, R.H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R.E. Petty & J.A. Krosnick (Eds.), Attitude strength: Antecedents and consequences (pp. 247–282). Mahwah, NJ: Erlbaum.

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. Current Directions in Psychological Science, 28(4), 331–336. https://doi-org.proxy.library.cornell.edu/10.1177/0963721419835206

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. Journal of Personality and Social Psychology, 38(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation : An integrative review of implicit and explicit attitude change. Psychological Bulletin, 132(5), 692–731.

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. Journal of Experimental Social Psychology, 40(4), 535–542. https://doi.org/10.1016/j.jesp.2003.10.005

Gladwell, M. (2005). Blink: The power of thinking without thinking. Little, Brown and Co.

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.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90,* 1–20. http://dx.doi.org/10.1037/ 0022-3514.90.1.1

Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain : The negativity bias in evaluative categorizations. Journal of Personality and Social Psychology, 75(4), 887–900.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6), 1420–1436. https://doi-org.proxy.library.cornell.edu/10.1037/0278-7393.20.6.1420

Kassin, S. M., Bogart, D., & Kerner, J. (2012). Confessions That Corrupt: Evidence From the DNA Exoneration Case Files. Psychological Science, 23(1), 41–45. https://doi.org/10.1177/0956797611422918

Levey, A. B., & Martin, I. (1975). Classical conditioning of human 'evaluative' responses. Behaviour Research and Therapy, 13(4), 221–226. https://doi.org/10.1016/0005-7967(75)90026-1

Levitin, D. J. (2014). The organized mind: thinking straight in the age of information overload. New York, N.Y.: Dutton.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior Research Methods, 47(4), 1122. https://doi.org/10.3758/s13428-014-0532-5

Mann, T. C., Ferguson, M. J., & Smith, E. R. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. Journal of Personality and Social Psychology, 108(6), 823.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. Behavioral and Brain Sciences (Print), 32(2), 183–198.

Munro, E. (1995) The power of first impressions, Practice, 7:3, 59-65, DOI: 10.1080/09503159508411629

Nyhan, B., & Reifler, J. (2010). When corrections fail: the persistence of political misperceptions [in the US]. Political Behavior, 32(2), 303–330.

Peters, K. R., & Gawronski, B. (2011). Are We Puppets on a String? Comparing the Impact of Contingency and Validity on Implicit and Explicit Evaluations. Personality and Social Psychology Bulletin, 37(4), 557–569. https://doi.org/10.1177/0146167211400423

Petty, R. E., Tormala, Z. L., Brinol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change : An exploration of the PAST model. Journal of Personality and Social Psychology, 90(1), 21–41.

Porter, S., Ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: the impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. Psychology, Crime & Law, 16(6), 477–491.

Rabin, M. & Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, *114*(1), 37–82.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 5(4), 296–320.

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. Journal of Personality and Social Psychology, 91(6), 995–1008. https://doi-org.proxy.library.cornell.edu/10.1037/0022-3514.91.6.995

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. Psychological Bulletin, 105(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3– 22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, E. R., & Decoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and Social Psychology Review, 4(2), 108–131.

Thorson, E. (2015). Belief Echoes: The Persistent Effects of Corrected Misinformation. Political Communication,33(3), 460-480. doi:10.1080/10584609.2015.1102187

Todorov, A. (2017). Face Value: The Irresistible Influence of First Impressions. United Kingdom: Princeton University Press.

Todorov, A., Mende-Siedlecki, P., & Dotsch, R. (2013). Social judgments from faces. Current Opinion in Neurobiology, 23(3), 373–380. http://doi.org/https://doi.org/10.1016/j.conb.2012.12.010

Victorino, L., Pilati, R., & Linhares, A. (2019). Priming and Prejudice: The Bias Effect of Origin Information on Peer Review, Judgment and Evaluation. Avances En Psicología Latinoamericana, 37(1), 169–178. https://doi-org.proxy.library.cornell.edu/10.12804/revistas.urosario.edu.co/apl/a.5635

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. Psychological Review, 107(1), 101–126.

Winter, L., Uleman, J. S., & Cunniff, C. (1985). How automatic are social judgments? Journal of Personality and Social Psychology, 49(4), 904–917. https://doi.org/10.1037/0022-3514.49.4.904 (Retraction published 1970, Journal of Experimental Psychology, 86[2], 255-262)

Wyer, N. A. (2010). You Never Get a Second Chance to Make a First (Implicit) Impression: The Role of Elaboration in the Formation and Revision of Implicit Impressions. *Social Cognition*, *28*(1), 1–19.