2020

Implicit Bias through the Lens of Electroencephalography

Hope Cooper

Implicit Bias through the Lens of Electroencephalography

A Thesis Presented

By

Hope Cooper

To

The Department of Psychology

In partial fulfillment of the requirements

For the degree of

Bachelor of Arts in Behavioral Neuroscience

Connecticut College

New London, Connecticut

December 2, 2019

Acknowledgements

This project was created with the help of so many mentors and family members to whom I could never get across the amount of gratitude that I feel.

First, thank you Dr. Jason A. Nier for your inspirational mentoring. Professor Nier both supported and pushed me in the exact moments that I needed it. He has taken an incredible amount of time to help me through this process and I cannot thank him enough. I sincerely appreciate all of the inspiration he gave me through this process which will hopefully lead me to a career in social science after Connecticut College!

I also owe an enormous amount of gratitude to Dr. Jeff Moher for his mentoring. I truly felt that he believed in me during every single step especially when I doubted myself. It is because of him that I was able to learn invaluable skills like how to set up and understand the EEG. Professor Moher is another extremely inspiring teacher and I will forever be grateful to have had the opportunity to learn from him throughout this process.

I must also thank Dr. Terry-Ann Craigie for stepping in to be a reader for me when I really needed her.

It has not always been easy to find such brilliant, passionate and inspirational professors like the ones mentioned above. I am so lucky to have found three professors to guide me through this intensive process so that I could learn from their brilliance and show me what I am capable of.

Finally, I would like to thank my parents for being incredible role models and encouraging me throughout my life, especially during this process.

Abstract

Unconscious or implicit bias is a part of everyday life. All human beings both exhibit implicit bias and (some more than others) are also the victims of it. Due to the way humans have evolved implicit bias will never be something that ceases to exist. Thus, it is important that neuroscience and social science closely study how it works and how to curb the behaviors caused by implicit bias. In the following research EEG (electroencephalography) was used alongside a weapons IAT (Implicit Association Test) to examine specific neural components that may correlate with higher bias scores on the IAT. Specific components N200 and P200 were examined. The results indicated that white and Black faces elicited different mean amplitudes in the N200 waveform. There was also a significant negative correlation between the difference of congruent (Black faces and weapons) and incongruent (white faces and weapons) groups and IAT scores (calculated as the D score). This result indicated that the smaller the difference between block types (congruent and incongruent) the larger the D score (bias). These results were the opposite of original hypothesis. They both support and extend the findings of previous research regarding implicit bias and EEG. Some results of the current study also give rise to new ideas about bias and cognitive control.

Table of Contents

List of Tables and Figures

List of Appendices

Implicit Bias through the Lens of Electroencephalogram

"I am not racist" is a sentiment too often and too commonly used by many white people when trying to defend or disguise systematic oppression and inequalities that continue to plague our societies. Many white people both in power, and not in power, try to reassure themselves by attempting to unconsciously hide the obvious ways in which white privilege and implicit bias are a part of our systems and behaviors. White privilege has been defined as the inherent advantages possessed by someone with white skin in a society that historically, and presently, gives privileged to people with white skin (McIntosh, 1988). Implicit biases are defined as unconsciously held set of associations about a specific group (Berghoef, 2019). Phrases like "I am not racist" or "I have Black friends" are the epitome of the ways in which implicit bias distorts reality. Many people who say phrases like those above understand a narrow definition of the word racist (because many dictionary definitions define a racist as anyone who benefits from white privilege). However, a broader issue is that many people fail to acknowledge a subconscious neural process, which results in the formation of in-groups and out-groups and the psychology and neurology behind it. These processes result in the ability to believe sentiments like the ones stated above without acknowledging the role that one's implicit bias may play in an already explicitly biased society. Understanding these underlying neural processes will bring to light a human attribute which will not be changed unless first examined closely and scientifically.

Classifying the surrounding world into categories is not only important from a modern functionality perspective but also from an evolutionary one as well (Brown, 2010). Categorization is quite literally how our brains function and allow us to get through our daily lives without being in danger (Brown, 2010). Literature on this topic suggests that stimuli are

sorted on the basis of similarity, which allows for the efficient processing of overwhelming amounts of input into the brain every second (Wang et. al, 2010). This constant stimulus intake is sorted into categories in order to simplify our perceptual worlds and allow us to be able to process information faster. Our brains purposely discriminate among stimuli in order to allow us to make judgements quickly and decide what the brain should attend to. This attribute works extremely well in the context of survival and safety. For example, if our brains had to choose between a lion and a butterfly they would likely tell us to focus more closely on the lion before the butterfly in order to preserve our safety. In this situation, if it was not innate for our brains to separately categorize the lion differently from the butterfly humans may have gone extinct because they would not have understood that the lion was in a category to be feared (Brown, 2010).

The neurological process of categorization becomes problematic in the context of 'sorting' fellow humans. Human beings are often sorted into categories such as race and gender, or what is familiar and unfamiliar to our previously defined categories (Brown, 2010). The individual whose brain is doing the sorting will often inflict their own bias, or over-arching cultural definitions onto the person in front of them (Wang et. al., 2010). Unfortunately, while in the process of sorting the brain also uses learned information in order to decide which category to place something in. For example, in the media golden retrievers are most often shown as loving family oriented dogs, however, pit bulls are often portrayed as aggressive fighting dogs. It is already known in our society that pit bulls are less likely to find homes, and are more frowned upon, not because this idea is based in reality, but because of a learned and preconceived notion that pit bulls are more aggressive. The brain categorizes pit bulls and golden retrievers differently and into familiar and unfamiliar categories. Golden retrievers are familiar and

although humans do not belong to this group Golden retrievers in the context of this study can be considered in the "in" or "congruent" group because of their familiarity and the absence of a learned threat, however pit-bulls which are less familiar and considered more of a threat are considered "out" or "incongruent". Similar categories work in the in the context of this particular study, due to a societal understanding of "threat" and lack of familiarity which makes humans implicitly biased towards the more threatening group also known as the out-group. Perceived threat even if accurate or inaccurate causes behavior to change drastically (Lundberg et, al,. 2018). The same process occurs when humans see other humans and their brains begin to categorize them into in-groups and out-groups (Lundberg et al., 2018). If a person is unknown, does not to fit into a category, or is wrongfully portrayed throughout history and present media, this person is subconsciously categorized in the out-group and not always but often, labeled as a threat (Lundberg et al,. 2018). The perception of threat is natural and is an important survival skill. However, it is easy to see how the brain, in combination with learned social expectations, can confuse these categorizations and can quickly create both overt and covert prejudice. As society continues to develop, humans have recognized that overt prejudice and discrimination are negative (not all people, but most). Although an individual might be aware of the detrimental nature of obvious discrimination and dehumanization (for example; slavery or implicit bias), this same person maybe unware of their implicit biases which may still be pervasive (for example; changing vocabulary to more simple words and phrases around someone who is perceived as poor or less educated). Implicit bias often occurs due to the process by which our brains subconsciously decode our surroundings in partnership with learned societal norms (Wang et. al., 2010).

The social psychology literature examining racial bias has been developing for over 60 years (since at least the early 1950's) and much research has followed such as the "Blue-Eyes Brown Eyes" experiment done in the 1960's (Peck, 2013). Even more recently, a large scale literature review conducted by Fitzgerald and Hurst (2017) examined a set of peer reviewed articles published in the span of ten years (2003-2013) in the field of medicine. The researchers wondered if trained healthcare professionals displayed implicit biases towards certain types of patients. After screening almost 300 articles (beginning with 4,000 possibilities, then researchers narrowed it down to the best 300 that answered their research questions) researchers concluded that healthcare professions did exhibit the same levels of implicit bias as the wider population (Fitzgerald & Hurst, 2017). Correlational evidence also indicated that biases are very likely to influence many different types diagnosis and treatment decisions especially in the heart disease categories (Green et al., 2007).

As one can imagine biases can heavily influence important outcomes which is why it is important to study and understand them in the brain. The effects of in-groups and out-groups, which have been documented by social science research are fairly well known (Peck, 2013). However, brain activity and the neural mechanisms at play are less well known. More recently, a literature review conducted by Stanley (2008) examined studies using fMRI, which measures BOLD (blood oxygen level dependent) to see brain activity in the context of in-groups and out-groups, and even implicit racial bias (Stanley et al., 2008). This literature review showed the brain areas that are most active in implicit bias, or during the formation of in-groups and out-groups are the amygdala, anterior cingulate cortex (ACC), dorsolateral prefrontal cortex (DLPFC) and fusiform face area (FFA) (Peck, 2013; Stanley et al., 2008). Once groups are formed and social categorization is achieved, research shows activation in insula and anterior

cingulate cortex, medial prefrontal cortex, and top down regulation in the dorsal lateral and prefrontal cortex, in order to mediate empathy (Kubota, 2012). These areas of the brain are most active during the process of determining in-group/outgroup membership, social categorization, action perception, empathy and, lastly, facial perception (Stanley et al., 2008). Together these processes often result in in-group and out-group biases.

Behavioral examples of implicit bias

Our implicit biases often influence the way in which we treat those around us and depending on the bias, our actions and behaviors can become implicitly unfair. For example, a study by Wang et al. (2017) focused on the relationship between bias and fairness by looking at asset distribution and brain activity. In this study participants were recipients in an "ultimatum game." The ultimatum game is a commonly used scenario in research where the proposer offers a division of the assets (given in the game) between themselves and another player (Wang et al., 2017). If the other player accepts the deal then both participants have more assets, however if the other player does not accept then neither participant gains any assets. Usually this is done in a double blind setting in order to show that people are often self-interested and rational (Wang et al., 2017). However, in Wang's (2017) study the exercise was not double blind and the goal was to examine the offers proposed to different recipients and the recipients' standards for fair and unfair. The rejection of an offer means punishing both the proposer and the recipient and the decision to reject or accept is telling about the recipient's ideas of fairness. Many social factors play into this decision. The behavioral results showed that acceptance rates rose in in-group partners rather than out-group partners. This result indicated that participants were more likely to see offers as unfair and reject them in an out-group partnership (Wang et. al,, 2017). This study also examined ERP (Event Related Potential) during the task. Specifically, they analyzed P3 and

FRN or feedback related negativity. FRN is reported to be most typically be observed as a negative deflection when a participant is responding to unfavorable outcomes compared to favorable outcomes. It peaks between 250 and 350ms. P3 is found to peak between 250 and 500ms and is also related to outcome evaluation. The results indicated that in-group and out-groups have very different effects on early and later processes of fairness shown by P3 and FRN waveforms. FRN amplitude was higher when offers were made by an in-group proposer (this outcome was significant) and when offers where made by an out-group proposer the effect was not statistically significant, meaning that the brain only seemed to be responding to in-group offers (Wang et al., 2017). Statistically significant and similar results for the P3 component occurred as well (Wang et al., 2017). This study is an important example of how social in-groups and out-groups can affect standards of fairness. We see examples of this frequently in everyday life, from micro-aggressions to police brutality and even in the field of medicine and science (Green et al., 2007).

While many studies use fMRI, an extremely important study (Wang et al., 2017) using EEG (electroencephalogram, which measures electrical activity in the brain) found that drastic changes in frequencies of electrical activity were modulated by in-group and out-group membership. This study showed that participants were more likely to accept unfair "offers" from members of in-groups rather than out-groups. This was consistent with the EEG measurements showing theta power (4–6Hz, low frequency waves) was larger for more unfair offers then fair offers in in-group interactions and showed no difference in out-group interactions. Measurement of mid-frontal theta activity was larger for negative feedback meaning that negative theta feedback was indicative of implicit in-group bias and theta oscillations were thought to be generated in the anterior cingulate cortex, as mentioned above in previous literature. Wang

(2017) concluded that due to electrical activity perceived ideas of fairness are largely modulated by in-group and out-group bias.

Another study (Green et al., 2007) examined doctors' diagnosis of thrombosis in Black and white patients in comparison with the doctors' IAT (the Implicit Association Test, commonly used to test implicit bias) scores. Doctors' IAT scores were directly correlated with their hesitation or lack thereof to treat Black patients. The same doctors who were less likely to diagnose and treat Black patients accurately also reported describing those patients of color as unwilling to listen and disrespectful. The same brain regions were analyzed via fMRI in this study, and similar results were found (Green et al., 2007).

A study by Sheng et al. (2013) examined the in-group out-group phenomenon but in relation to the neuropeptide oxytocin. The purpose of the study was to link oxytocin to either in-group favoritism or out-group empathy (Sheng et al., 2013). The researchers tested their hypothesis on Asian males looking at the faces of Caucasian males which were expressing pain or showing a neutral face. Oxytocin was self-administered, or a placebo was self-administered. Instead of looking at specific brain regions like previous studies mentioned above this study used EEG/ERP to examine sizes of amplitudes relating to seeing pain or a neutral face (Sheng et al., 2013). The results suggested that self-administered oxytocin had the largest effect on in-group facial expression and modulated out-group expressions of pain. ERP data showed a greater peak amplitude at P2 for in-group neutral expressions and a lower peak amplitude for out-group neutral expressions but P2 in general had higher peak amplitude for expressions of pain suggesting that oxytocin has an effect in shaping empathetic neural response (Sheng,.et,al. 2013). These results show an important relationship between in-group and out-group relations and how

they can be varied slightly by a chemical response in the brain. ERP and EEG studies and effects will be discussed more thoroughly later in the paper.

In-groups and out-groups, congruent and incongruent

First and foremost, it is important to examine what occurs when social categorization goes wrong and creates harmful prejudice. It was argued by Bruner (1957, cited in Brown 2010) that categorization of in-groups and out-groups is an inescapable quality of human existence. Categorization at the simplest levels allows humans to discern different parts of their surroundings. However, what is most perplexing about this natural human phenomenon is what occurs when a simple evolutionary trait of convenience becomes a matter of life or death (Brown, 2010). There are many examples similar to the butterfly and lion analogy mentioned above in which social categorization is important for survival. In fact, even the categorization of fellow humans can be important for safety. Brown (2010) writes that it is important to know how to recognize "us" or "them" categories in certain places in the world. One place he discusses is Jerusalem, where if you went for a walk alone on the street it would be important in this modern time to be able to discern who and where is safe and what or who to not approach. (While this is true, it also seems important to think about the idea that if humans could put their ability to categorize their social surroundings aside it is possible we would not have conflicts like the ones in the Middle East in the first place). However, in the process of understanding why our brains categorize in the way they do, it is vital to examine where this process of categorization originates and what it does to our cognitive functioning in the moment. Brown (2010) discusses these origins, and the changes in cognitive functioning that may allow bias to persist after the initial categorization.

A series of well-known intergroup experiments were conducted by Tajfel and colleagues (1971, cited in Brown 2010). These experiments examined the idea that the origins of bias begin when the perpetrator of the bias first feels a sense of belonging to their in-group rather than when they first see an out-group. In society we often think of this when talking about white privilege. White privilege is described as an intrinsic part of whiteness. When someone is born with white skin they automatically benefit from their whiteness. This phenomenon is best described in the famous essay by Peggy McIntosh, *White Privilege: Unpacking the Invisible Knapsack*, where McIntosh writes about white privilege as an unseen unconscious advantage that individual born into whiteness benefit from (McIntosh, 1989).

For example, skin color when being pulled over by the police in certain parts of the United States can be the difference between life and death. Just the possibility of being pulled of increases by 30% if the person driving is Black (Stanford Open Policing Project, 2019). This is an example of how white privilege works. There is a connection between previously discussed literature on white privilege and what Tajfel and other researchers argue in their hypotheses about the origins of bias. Bias originates, like white privilege from belonging to a group. This is known as the minimal group paradigm (Brown, 2010). The creation of the phrase "minimal group paradigm" came about through a series of experiments in which scientist sought to create a model of group structure without all of the normal attributes that usually create group dynamic. This included taking away group life, face-to-face interaction, a set of norms, and relationships with other groups. The only thing that made participants feel as though they were in the group was the act of being told that they were a part of a specific group and not a part of another specific group. This set of studies was called the Klee and Kandinsky experiments named after the artists of the paintings used in the experiments (Tajfel et al., 1971, cited in Brown 2010). In

one of these studies participants took part in a decision-making experiment. Participants were asked to choose from pairs of abstract paintings and indicate which paintings they preferred, then, on the basis of these choices participants were assigned to one of two groups. The next part of the experiment involved the distribution of money to other people in the study who were only identified by using numbers and group membership. Prepared booklets of reward matrices were given to participants, and participants were not allowed to award money to themselves. Also, participants had to choose one pair of numbers which stood for real monetary rewards. Matrix one was designed to measure general in-group favoritism and matrix two was designed to measure the tendency for participants to make the difference between in-groups and outgroups even more spread apart. The table below is adapted from Tables 2 and 7 in Tajfel et al.,1971 and shows two sample matrices from the minimal group paradigm.

Table 1.

Reward Matrices -This table was adapted from Brown 2010, wherein Brown describes research conducted by Tajfel (1971)

Matrix	Reward points
Matrix 1 Member 72 of Klee Group Member47 of Kandinsky group	 18 17 16 15 14 13 12 11 10 9 8 7 6 5 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Matrix 2 Member 74 of Klee Group Member 44 of Kandinsky Group	 25 23 21 19 17 15 13 11 9 7 5 3 1 19 18 17 16 15 14 13 12 1 10 9 8 7

In the experiment each matrix was presented to each participant twice, once similar to how it is listed in Table 1, and once with the group affiliations of the recipients reversed. Participants did not know who members 72 and 42 were, yet their job was to allocate the money in the fairest way. Researchers concluded that the most common response was for participants to attempt to be fair but at the same time show a reliable tendency towards awarding more money to in-group members then to outgroup members. In matrix 1 over 70 percent of participants made choices that favored their group with a mean response from people in the Klee group of somewhere between 14/9 and 13/10 (money is represented in a ratio, 14/9 would indicate $14 to the in-group and $9 to the outgroup). Similar Choices were made in matrix 2 even when it meant that the in-group may not be as "rich" as possible participants chose the numbers that would insure that they had more money than the other group. For example, Brown writes that in matrix 2 the mean response from those in the Kandinsky group was somewhere between the 13/13 and 12/11 choices (Tajfel et.al.,197, cited in Brown, 2010). This choice, however, results in the Kandinsky in-group receiving 6 or 7 points less than they could have received. The Kandinsky in-group does receive more than the Klee group which is what researchers hypothesized would be the main goal of the participants. This strategy of making sure that each participants in-group receives more points or money than the outgroup while not taking into consideration the best possible outcome for their own in-group is called the maximizing difference strategy. Researchers set up this study in a way that participants would feel no connection to their in-group other than the title of the group itself and still this caused a bias in most participants. This showed that just the very act of belonging to a group caused participants to act differently and thus show a change in behavioral prejudice during the task. Brown (2010) also suggests that this

study demonstrates how bias created cognitive changes that perpetuate even more bias and make an endless circle.

Brown (2010) also describes a study conducted by Campbell (1956, cited in Brown, 2010) that is a clear example of how bias changes perception and cognition. In this this study, participants were asked to learn the physical location of two types of nonsense syllables. In each trial the given syllable was always presented in the same position. In group one the syllable always had the middle letter E and in the second group the syllable always ended with X. The E group was always to the left and the X group to the right however in the middle of the physical space the syllables overlapped. Participants were consistently wrong in estimating the placing of the overlapping syllables (Campbell et al., 1956, cited in Brown, 2010). Participants typically exaggerated how far over the syllable must be. "E" participants guessed extremely to the left, while "X" participants guessed extremely to the right. Hypotheses describing these exaggerations may occur in the human psyche were later written about by Tajfel and Wilkes (1963, cited in Brown, 2010) Tajfel and Wilkes (1963) claimed that if a scenario is set up to impose a distinction between types of categories like group "A" and group "B" or "Blue" versus "White," be it physical objects, sensory events, or even people, participants pre-existing awareness of differences between the two groups will be enhanced. Participants perceived different groups as even more different than they actually were and saw similar groups as even more the same then they were in reality. These ideas show that cognitive function changes when a stereotype exists and allows this allows for a subject to act on their bias in an even more extreme way (Tajfel & Wilkes et al., 1963, cited in Brown, 2010).

Another study looked at the behavioral consequences of the above hypotheses. The study was conducted by Correll et al. (2002) at the university of Colorado at Boulder. In this study,

researchers examined the behaviors of participants and how ethnicity might affect their decision to shoot or not shoot characters in a video game. The study as a whole consisted of four small parts all looking at the same main idea. Study one focused on different suspects in the video game and their ethnicity, and researchers found that participants fired on an armed target faster when the target was Black then when the target was white. Participants also decided to withhold shooting more quickly when the target was white then when the target was Black (Correll et al., 2002). In study two, the same study was repeated with an emphasis on time restrictions. Participants in this study more often mistakenly shot an unarmed person when the target was Black. Researchers concluded that time did not change participants ability to see the difference between and unarmed and armed target depending on ethnicity, however, participants required less certainty in making their split second decision to shoot if the target was Black (Correll et al., 2002).

Study three was similar to the first study but researchers investigated the prior beliefs of the participants before beginning the same set up. A Modern Racism Scale questionnaire and the Discrimination (DIS) and Diversity scale (DIV) were added to the video game task. Behavioral results in this study showed no correlational link between scores on these questionnaires and being able to predict shooter bias. The researchers hypothesized that because the questionnaires were related to perceptions of cultural bias rather than personally endorsed stereotypes, that the mere knowledge that the stereotypes exist is enough to induce shooter bias. Lastly, in study four researchers found support for their hypothesis from study three by changing their participant pool to both black and white participants rather than only white participants. Both black and white participants displayed the same type and level of shooter bias, suggesting that just the predication of bias in society creates real shooter bias. The results from these four studies did

support the hypothesis that the decision to shoot or not to shoot is heavily informed by the ethnicity of the target and the bias of the shooter (Correll et al., 2002).

There are too many examples of how, both subconsciously and consciously, police officers may allow their bias to influence their cognitive sorting abilities. Just in 2019 alone more than 150 unarmed black citizens have been shot and killed by police in the U.S (washingtonpost.com, 2019). The shooting of unarmed Black men has often been attributed to fear and self-protection, however in this study researchers show that the decision to shoot correlates directly with the race of the target and that both white participants (studies 1-3) and a community of white and Black participants (study 4) are capable of making wrong and bias decision to shoot. This study exemplifies the ideas written about by Tajfel and Wilkes, discussed above, showing that differences between or within the two categories (Black and white) will often be attenuated and maximized (Tajfel & Wilkes et al., 1963 cited in Brown, 2010; Correll et al., 2002).

The shooter bias is not the only window that scientists have into looking at the origin and maximization of bias. Sadly, bias exist in almost every part of human life. Understanding how it occurs both behaviorally and neurologically could help begin the process of educating people to curb the biases that negatively impact both themselves and the people around them.

In the context of curbing implicit bias, it is important to examine the social, behavioral and neurological processes that contribute to in-group and out-group formation. While there are many social science models to help researchers understand the human processes involved in bias, brain imaging is also important for understanding this process of categorization and the issues that unfold because of it. Particularly, EEG allows researchers to have a neurological and time

sensitive understanding of how brain activity changes after participants are presented with either in-group or outgroup stimuli.

EEG and ERP

Using EEG/ERP is extremely important in examining the time sensitive reactions that occur during instances of bias and prejudice. ERP (Event Related Potential) has greater temporal resolution then other forms of neural imaging, and this is important for understanding how information is processed rather than just what the brain does with the information (Payne, 2006).

EEG is extremely informative and useful in determining changes in the brain during in-group and out-group experiences (Luck, 2005). Studies using EEG and ERP have been conducted to understand brain activity and implicit bias. ERP or time-locked EEG is a much clearer way to interpret EEG data. Raw EEG data encompasses all activity going on in the brain, or at least all that the skullcap of the EEG captures. Raw EEG data, before it is computed into ERPs, also shows facial muscle contractions, blinking, or anything that is referred to as "noise." This is because the EEG cap electrodes are so sensitive they also pick up on electric stimuli that is not related to the brain. By using ERP most of this extra noise is accounted for and computed out when the ERP results are calculated. Extracted ERP data allows researchers to see changes in ERP waveforms that solely have to do with the stimulus at hand. (i.e. ERP cancels out "noise" from brain activity that is not due to the stimulus that is part of the study; Luck, 2005; Sur, 2009). ERPs are elicited by many different events; sensory, cognitive, or motor (Sur, 2009). ERPs occur when "similarly oriented neurons" fire by the thousands or millions in synchrony due to an internal or external stimuli or event. This firing of neurons takes place after stimulus occurs, and is a symbol of the brain's reaction and processing of information (Sur, 2009). Waveforms or components of ERP can be understood as either sensory/exogenous, and these

include components that peak within the first 100 milliseconds after the stimulus occurs, or cognitive or endogenous ERPS which occur later and signify processing of information (Sur, 2009). In the current study both types of components will be examined.

EEG and ERP data has been extremely useful in determining differences in neurological response to stimuli in children as early as 7 months old. One study conducted by Elsner et al. (2013) examined infants in a three-stimulus oddball paradigm (a sequence of repetitive stimuli that is interrupted by unexpected stimuli). In this study there were two conditions: one with an animate object picture as the standard stimuli and an inanimate as the oddball or contrasting stimuli, and the second condition with an inanimate object as the standard and animate object as the oddball-different or contrasting stimuli. Discrimination of oddball or contrasting stimulus was apparent in the infants ERP results. This difference was most clear in "Nc peak amplitude and late-slow-wave activity" for the contrasting stimuli (Elsner, Jeschonek & Pauen, 2013). Larger Nc latency and positive slow-wave activity showed greater processing for the contrasting category compared to the same category. ERP patterns showed similar results in oddball same stimulus rather than oddball different stimulus, (i.e. chair, dresser or giraffe, bunny). The researchers concluded due the ERP component patterns that the 7-month olds in the study decoded stimuli by perceptual features and also by category membership. This study showed significant differences not only between similar and different stimuli but could also show differences between living and non-living stimuli. Future research may be able to connect waveform patterns in this study to patterns in EEG studies regard in-group and out-group stimuli and possibly help address these biases at a young age (Elsner, Jeschonek, Pauen, 2013).

Another study conducted by Wang et al. (2010) examined ERP data of participants when asked to differentiate between semantic stimuli of either in-group or out-group people. In this

study, 16 young subjects were asked to classify adjectives as either positive or negative while their ERP data was recorded (Wang et.al, 2010). The study aimed to determine what ERP waveforms occur when a participant may be feeling prejudice or bias against a group, in this case the "out-group" was rural migrant workers (RMWs) in China. RMWs are a social group under the household registration system that was set up in 1958. This system separated urban and rural citizens and does not allow them to live in areas that they were not born in unless they possess a *hukou* which in the past has allowed citizens from rural areas to live temporarily in urban cities in order to find work. Typically, these RMWs perform jobs that the citizens of the city will not do. In Chinese society there is a clear bias against RMWs, they are considered crude, unintelligent and uncivilized (Wang et.al, 2010). When primed with stimuli associated with RMWs the reaction times to identify positive adjectives were longer than when primed with stimuli associated with urban workers. For example, "urban worker; clean" had a faster ERP reaction time then "RMW; clean". This result suggests that RMW and clean were incongruent concepts, and thus showed a bias against RMWs. Specific and significant components involved in this study will be discussed later in this paper, however the waveform pattern evoked during an RMW-positive condition versus RMW-negative was significantly larger suggesting that there is a greater conflict between the idea of RMWs and positive adjectives, against suggesting negative bias against rural migrant workers (Wang et.al, 2010).

A third study, which examined the relationship between race-related implicit associations and EEG evoked by faces from different races, found that ERPS for black and white faces differed significantly (He et al. 2009). Black faces evoked a larger positive ERP that peaked at 168 ms over the frontal scalp, and white faces evoked a larger negative ERP the peaked at 244ms. The results correlated significantly with participant's race IAT (He et al., 2009) This

study suggested that in-group stimuli may be consistent with larger negative amplitude ERP whereas outgroup stimuli may result in larger positive ERP. Other literature suggests that outgroup stimuli may coincide with larger amplitudes for both N1 and P2, whereas in-group stimuli may relate to smaller amplitudes in both N and P components (Campenella et al., 2002).

Thus overall, previous research suggests that specific ERP components are associated with different cognitive functions. Researchers can identify these specific wave forms and find patterns in order to understand some of the processes that underlie specific behaviors or information processing (Payne, 2006).

Specific Components

Different components are described as unique ERP waveforms that differ in latency and amplitude (Sur & Sinha, 2009). Components are referred to with letters N or P each letter indicates polarity within the waveform; negative and positive. Each letter is followed by a number indicating latency or position in the waveform (Luck, 2009). For example, the first negative peak often occurs 100 ms after the stimuli, and is usually referred to as N100, similarly the second peak is often a positive peak followed by a 100 or 200. Another category separating types of components is early and late. Early components are associated with immediate reactions to stimuli, whereas later components such as P300 which varies and may occur between 250-700 ms coincide with cognitive processing (Luck, 2009).

Peaks in certain components have been thought to signify specific attitudes or ways of processing in that moment. For example, a N100 peak has been thought to reflect early selective attention and discriminative processing rather than detection processing. N100 occurs between 150-190ms. Another example is P200, which often follows the N1 wave and occurs around 200 ms after the stimuli presentation (Volpert, 2012). This component is often larger when stimuli

contain a target feature and the amplitude will likely appear even larger when the targets are sporadic throughout the trial (Luck, 2009). P200 is also known to have larger amplitudes when threatening stimuli occur such as out-group faces and negative information (Volpert, 2012).

N400 is a negative-going ERP component occurring at 400ms after presentation of the stimuli. N400 is often linked with the integration of a semantic stimuli into a specific context (Volpert, 2012). N400 is also an indicator of the activation of a response from a primed concept and the difficulty of associating two concepts. Larger N400 should occur when incongruent stimuli are presented. For example, the pairing of "aggressive" with the word "women" has been shown to elicit a greater N400 then the pairing of the words "women" and "nurturing" (Volpert,2012; White et al., 2009)

As researched by Polich (2007) P300 has been shown to correlate with high implicit bias scores on an IAT task. P300 occurs at 330-380 ms and is observed in any task that requires stimulus discrimination. P300 is also associated with attention, memory and habits that are associated with targeting stimuli. P300 is also task specific meaning that more difficult or longer tasks often reduce P300 amplitude but lengthen peak latency (Volpert, 2012). This could be extremely helpful in analyzing implicit bias due to P300s nature of being present during the moment a participant needs to decide which group each image belongs to in an IAT. P3a is closely associated with memory processing and distractions as well (Polich, 2007).

One study conducted by Volpert (2012) examined whether explicit and implicit racial biases would moderate EEG activity. In this study researchers focused on components N100, P200, N400. Participants completed a face train pairing task in which positive and negative (stereotypically black or white traits) followed the presentation of either a black or white face. Results indicated that P200 was changed drastically by evaluative implicit bias where as N400

activity related more closely to self-reported explicit bias (Volpert, 2012).

Another study, which examined ERP and snap decisions to shoot characters in a video game, focused on ERP analysis on early P200 and N200 waveforms. In the study conducted by Correll et al, (2006) and as discussed by Payne (2006), participants were told to shoot only the characters in the video game that were holding weapons. Half of the characters were Black and half white. While in reality none of the characters were holding weapons, participants shot a greater number of Black characters than white. EEG/ERP data was recorded while participants played the game. N200 and P200 showed a significantly different change when the participants target was Black versus when the target was white. The same results occurred for waveforms when looking at armed or unarmed targets. A larger P200 amplitude and smaller N200 occurred for participants who had a higher rate of shooting unarmed Black characters and thus showed more bias during their task. The researchers speculated that these ERP patterns were likely due to the participants attempting to control the shooter bias rather than the bias itself. P200 was also associated with emotional reactions to threats or threatening stimuli while N200 peak appears during conflict detection and cognitive control. The study found that participants with greater P200 amplitude responses to black individuals, and those with lesser N200 responses, showed greater racial bias (Correll, Urland, & Ito, 2006; Payne, 2006).

Cognitive control is also important to take into account when considering how waveforms are effected by bias or by any type of stimuli. One study conducted by Kumar et al., (2010), examined the effects of acute moderate exercise on cognition for people with sedentary lifestyles. Researchers in this study wanted to examine the effects that exercise had on cognitive control on different wave form latencies. The researchers concluded that the waveforms most effected (where the latencies changed the most) by cognitive control were N200 and P200.

(Kumar et al., 2010). Another important study that dealt with cognitive control and EEG used a novel task with partially incongruent categorization in order to examine the timing and effects of cognitive control on ERP data (Chen et al., 2009). In this study participants needed to categorize stimuli by features that corresponded to a given specific concept. The concept and the number of features was varied. Researchers hypothesized that when there was more than one feature on the stimulus participants would have to elicit more cognitive control in order to decide what category the features responded to. The results of the study indicated that cognitive control elicited a higher peak N200 ERP and a higher P300 ERP in the anterior cingulate cortex and prefrontal cortex respectively. These results support the idea that N200 corresponds to conflict and thus also may create a higher P300 ERP amplitude in order to respond to the given conflict (Chen et al., 2009).

"Threat" can also be applied to in-group and out-group stimuli and can be expressed in ERP waveforms. Participants often find out-group stimuli more threatening. Other studies that look at in-group and out-group stimuli in relation to ERP have found similar patterns and have also noticed amplitude changes in N400 when focusing on out-group stimuli (Wang et al., 2010). One of the studies discussed above in the EGG/ERP section, conducted by Wang et al. (2010), is a good example of what the amplitude of the N400 component indicates. N400 is a late component meaning that it occurs during information processing. This is important because when taking in a visual stimulus, participants first see the stimulus and then form a judgement. When the judgement is made other factors come into play, such as ones' awareness of bias and trying to prevent the bias. The opposite might also occur with the judgement reflecting a feeling of threat or discussed with the out-group (Luck, 2009). N400 often represents a conflict in processing. In this study the amplitude of N400 during RMW–positive adjective condition was

significantly larger than in the Urban worker-positive adjective condition. These results suggest a greater conflict between positive adjectives and RMWs (Wang et al., 2010).

P3a, which usually peaks between 250-280 MS, is also an important waveform that peaks when novel stimuli are shown. P3a is often paired in research with N2, because peaks in amplitudes from both of these waveforms often occur with new or surprising stimuli and both waveforms are time-pressure sensitive, meaning they elicit greater amplitudes when participants need to make decisions within a certain amount of time.

A study conducted by Campanella et al. (2002) examined the emotional facial expression, using an oddball task to determine patterns of N2/P3a waveforms. The purpose of the study was to determine if complex visual stimuli, such an emotional expression, could bring about a peak in both N200 and P3a. If a greater amplitude did occur the findings from the experiment would indicate that subtle changes in facial expression could elicit similar ERP results to in-group/out-group stimuli. The findings of this research did show a peak in N2/P3a amplitude for deviant stimuli and low N2/P3a amplitude for frequent stimuli. The examination of the relationship between these two waveforms is important to any ERP research when novelty or fear is involved, and thus will be important to examine in the course of the current study (Campanella et al., 2002).

Table 2.

Relevant Components, their placing on the scalp and their time in MS after the stimulus is shown

N100	Around 100 MS discriminative processing Often occurring at: Fz, Cz, Pz, F3, F4, C3, C4, P3, P4, CP1, CP2, FC1, and FC2
N200	Around 200-280 MS Conflict detection Often occurring at: Fz, Cz, Pz, F3, F4, C3, C4, P3, P4, CP1, CP2, FC1, and FC2.
P200	Around 200-280 Target feature processing (looking for a specific target) Occurring at: anterior and central sites Often occurring at: Fz, Cz, F3, F4, C3, C4, FC1, and FC2
P300/ P3a	Around 300 -380 Novelty stimuli Often occurring at: Fz, F3, F4, FC1, FC2, Pz, P3, P4, C3, C4, Cz, CP1, CP2, CP5, and CP6
N400	After 400 MS Conflict monitoring Occurring at: Frontal-central:F3, FZ, F4, FC3, FCZ, FC4, CZ, C3, and C4
P150	Around 150 MS Early perceptual processing, social categorization Often occurring at: Fz, FCz, Cz, CPz and Pz

Implicit Association Test

The IAT (Implicit Association Test) is a measure of implicit bias in which strength of automatic associations can be calculated (Greenwald, Benaji & Nosek, 2003). IAT score can be determined by computing performance speeds during two performance tasks. The latency of association greatly depends on strength of association (Greenwald, Benaji & Nosek, 2003). Each IAT task is split into seven blocks. Five of the blocks are practice blocks where the subject must choose between two keys, usually "I" and "E", when different stimuli appear on the screen. Two of the blocks are called test blocks. The purpose of the test blocks is to isolate the categorical latencies. Some test blocks are congruent blocks and some are incongruent blocks in order to test the strength of association between categories. For example, Black faces and weapons or violence would be used as a congruent group due to the familiarity and ingraining of this stereotype in our minds. In contrast, white faces and weapons would be incongruent due to the lack of familiarity in that we most often see Black faces and weapons rather than white faces and weapons (Lundberg et al., 2018). The specific procedure employed in the current study will be discussed later.

Other studies use congruent and incongruent stimuli in-groups and out-groups. For example, one of the first papers published on the IAT by Greenwald, Benaji and Nosek (2003) uses four categories to exemplify a basic IAT measure. Researchers use images of George Bush, Al Gore and pleasant and unpleasant words. The goal of this procedure was to understand what types of words (pleasant or unpleasant) are associated with each politician. In order to effectively measure bias, the latency between seeing an image and pressing the corresponding key was measured. In some blocks Al Gore and pleasant words were paired at the same key and in others unpleasant words and Al Gore are paired. The same was true for images of George

Bush. The set-up of these blocks successfully produced typical IAT results and exemplifies how following IATs should be set up.

Table 3.

Block Procedure as adapted from "Understanding and Using the Implicit Association Test: An Improved Scoring Algorithm" by Greenwald, Benaji and Nosek (2003). This set up of the IAT used the example of Al Gore and George Bush

BLOCK	# OF TRIALS	FUNCTION	ITEMS ASSIGNED TO LEFT KEY RESPONCE	ITEMS ASSINED TO RIGHT-KEY RESPONSE
1	20	Practice	George Bush images	Al gore images
2	20	Practice	Pleasant words	Unpleasant words
3	20	Practice	George Bush +Pleasant Words	Unpleasant words+ Gore images
4	40	Test	George Bush +Pleasant Words	Unpleasant words+ Gore images
5	20	Practice	Al gore images	George Bush images
6	20	Practice	Pleasant words +Gore images	Unpleasant words+ Bush images
8	40	Test	Pleasant words +Gore images	Unpleasant words + Bush Images

Healy, Boran and Smeaton (2015) describe the IAT as "a reaction time based categorization task that measures the differential associative strength between bipolar targets and evaluative attribute concepts as an approach to indexing implicit beliefs or biases" (Healy, Boran & Smeaton, 2015, p.1). The IAT effect or score is the calculated standardizes difference (D) between the mean latencies of congruent (in-group) and incongruent (out-group) pairings. A positive D score indicates that a participant is slow to respond to incongruent pairs and fast to respond to congruent pairs and vice versa. In general, researchers indicate which pairings are congruent and which are incongruent when deciding which bias they intend focus upon. In the study conducted by Healy et al, (2015), researchers sought to understand the ERP waveforms that occurred during a general IAT task in order to further research how well the IAT and ERP measures work together in research. The IAT task in this study was purposefully constructed to be somewhat basic in order to keep the main focus on the ERP data that corresponds to the IAT task. The categories in the IAT were "nature" versus "built" or man-made and "me" versus "other". In this study "me-built" was congruent and "nature-other" was incongruent. An increase in reaction time in the incongruent blocks and a decrease in reaction time in the congruent blocks shows a pro-nature bias in this study. In the comparative ERP section of the study, researchers found no significant results in the N1 and P2 waveforms but did find that N2 amplitudes were more negative in correlation with high D scores. They also found that N2 was enhanced (more negative) during congruent blocks rather than incongruent blocks. P300 ERP amplitudes were higher in high to medium D score groups then in low D score groups. In short N2 and P300 were enhanced in correlation with high D scores. This meant that the more bias a persons' IAT effect appeared (the higher the D score), the more drastic the N2 and P300 amplitudes appeared (in either direction, positive or negative). These findings suggest that EEG is important to

understanding the neurological processes that occurs during the IAT.

The same IAT procedure and scoring algorithms based on latency have been used in many other studies that examine different versions of the IAT. One such study used ERP data and the IAT to examine the difference in the results when the IAT measure itself was framed as a test of morality versus a test of competence (Nunspeet et al., 2014). In this study conducted by Nunspeet et al, (2014). the IAT was structured to present photos of Muslim women with hijabs, non-Muslim women and then chosen photos that were either "positive" or "negative" images. Each group showed a strong bias towards categorizing negative photos and Muslim women together, however, as expected, the IAT effect (the bias score) was weaker for those participants that were told the IAT measure was a test of morality, relative to those who were told the measure was a test of competence. This suggests that when the test itself was categorized as a moral test participants showed some control over their attention and ultimately their bias (Nunspeet, et,al,. 2014). Researchers also calculated ERP data and compared those data to the IAT results. Overall N1 was larger for out-group stimuli and smaller for in-group stimuli. The researchers also found that N1 was more significantly different for in-group and out-group stimuli in the morality conditions and not significantly different in the competence conditions. P150 was also larger for out-group stimuli than in-group stimuli, and the difference between in-group and outgroup stimuli was more pronounced in the morality group for p150 as well. N2 had a bigger amplitude associated with in-group stimuli but had no significant difference between groups. The researchers concluded that morality drives explicit behaviors and that IAT scores change depending on procedure and the type of measured bias. This study is also a helpful example of how ERP and IAT analysis paired together are extremely important in understanding the processes of bias in the human brain (Nunspeet et al., 2014).

The Weapons Bias

Stereotypes portraying Black men and women as hostile, aggressive, and often violent plague our culture today. There are countless examples of how these stereotypes have affected, politicians, police officers and regular citizens in ways that are destructive not only the Black community, but to all whom participate in being affected by or perpetrating those stereotypes (Lundberg et al., 2018). As cited in Lundberg et al (2018), on July 16 2009, a Harvard University professor, Henry Louis Gates, came home to find the lock on his front door broken. He, like any regular person, went to the back of his house in order to enter through the back door. During the time it took him to do that his neighbors had already called the police who would come to his property to arrest and charge him with disorderly conduct. Gates who was 59 at the time and walked using a cane, was unlikely to be mistaken for a criminal. However, in the United States, Black men are perpetually associated with violence and criminality due to the hundreds of years of slavery and racism that still have effects on our systems (Lundberg et al., 2018). While this topic as a whole would be much better suited for another paper due to its complexity and length, the story of Professor Gates is a relatively mild example of how horribly the bias of communities and police officers can affect fellow human beings.

Measuring implicit bias in the lab can help to understand the contexts in which it occurs, and, if possible, how to prevent it. There are many ways to measure implicit bias, one of which is the IAT. There are many types of IATs that test different biases. One version of the IAT is the Weapons Bias IAT test, which tests participants' familiarity and bias to seeing both Black faces paired with weapons and white faces paired with weapons. In the weapons IAT the groups are "Black faces, "White Faces", "non-harmful objects" and "weapons". There are many studies that have already used this version of the IAT, or similar, tasks to show patterns in implicit bias.

One of the most obvious and violent ways that the stereotyping of people of color manifests itself is in police violence and police shootings. In a split second an officer must decide how to go about their job. As we have learned already from studies mentioned above, time constraints make bias even more prevalent. These time pressures can result in police officers seeing weapons or sensing danger when there is none (Payne, 2006). One study, conducted by Payne (2006) used a task very similar to a version of the IAT that deals with weapons bias. Payne created a task where images of faces were flashed on the screen one at a time, sometimes white, and sometimes Black. After the quick image of the face an image of either a non-harmful object or a weapon appeared (this is different than the IAT because participants did not need to identify the race of the person who was in the image). Participants responded only to the images of objects that flashed on the screen after the face. Participants needed to identify if the object was a weapon or if the object was harmless. In one version of the study researchers asked participants to respond at their own pace. In another version of the study participants needed to respond to the image within half a second of the presentation of the object. The study showed that in the self-paced trials, participants were very accurate regardless of race. However, in the timed trials participants falsely detected a weapon much more frequently in the presence of a Black face. Payne, suggested that this occurred because the Black face primed the participants to detect a gun (Payne 2006).

In a similar study, researchers examined if intentional awareness of race, and racial bias, would change the pattern of bias observed (Payne, Lambert, & Jacoby, 2002). This study separated participants into a baseline condition where participants completed the weapons task but were told to ignore the faces all together. A second group of participants were told about the possibility that the faces might prime them to have a specific bias and to be careful not to

succumb to their implicit bias. A third group was told to intentionally use the race of the faces to help guess if the object in the next image was a weapon or not. The results showed that the participants directions for the task significantly affected their self-reported intentions and ideas surrounding the study in general. However, the participants' performance, or bias score, in the task did not vary at all. Racial bias emerged in all three groups and was even more noticeable in the "avoid bias" and "use bias" groupings. Researchers suggested that the mere implication of directing the participants to focus on race backfired on their performance and that a weapons bias seems to be separate from the intent of the participant. This indicated that the bias was likely to be truly implicit and that implicit bias can even exist when individuals are making a conscious attempt to be fair (Payne, 2006; Payne, Lambert, & Jacoby, 2002).

Split second decisions, such as deciding to shoot in a video game or in real life, often magnify the inaccuracies of the brains ability to intake and sort data. Thus, time pressure intensifies the impact of implicit bias of an individual. The weapons bias exemplifies this, and unfortunately mostly occurs most against Black men in the form of police brutality or killing unarmed Black men, which we too often see in the news (Lundberg et al., 2018).

Another related study examined the impact of the of time pressure by examining the ERP components associated with weapon biases (Correll, Urland, & Ito, 2006). The results of this study indicated that bias dictates how stereotypes will manifest in certain situations. In this study the stereotype that Black people are more linked to danger was evident in the participant's quick decisions to shoot unarmed Black video game characters more frequently than white ones. These stereotypes are often learned and they become biases that are quickly brought to the surface when making split second or subconscious decisions. These studies serve as a good example for why using the Weapons IAT in the following study is important; it highlights how biases become

more evident in split second decision making and also is a model for many other types of implicit bias that exist in our society (Correll, Urland, & Ito, 2006).

The Current Study

The current study sought to build upon the finding of previous research, and to explore novel hypotheses regarding the relationship between brain activity and implicit racial biases, and also to look closely at the neurological effects of the weapons bias.

Previous research has indicated a significant difference in EEG/ERP data for in-group and out-group stimuli (He et al., 2009; Sur et al., 2009). These studies, which used ERP, suggest that N1 or N100 (occurring around 90-200 ms) amplitude would indicate unfamiliar stimuli. Thus, a more biased person would likely have a high vertex potential (meaning a maximum amplitude over Cz). Therefore, a high N1/ P2 component may correlate with a more biased IAT result (He et al., 2009; Sur, et al. 2009). Previous research also shows N2b and N2C waveform may be present because of the nature of the IAT and the need to categorize the stimuli shown during the testing. N200 waveform maybe enhanced (more negative) for congruent (in-groups) groups rather than for incongruent (outgroup) groups.

P300 and P200 amplitudes may also be larger in high to mid scoring D scores in the IAT, meaning that the more bias a participant shows the higher likelihood of a larger ERP amplitude in p300. N400 may also show a higher amplitude in correlation with participants showing more bias. Lastly, a late positive component (around 592 ms) has been shown to be greater for familiar stimuli (He, et al., 2009).

The hypothesis of the current study was that more biased IAT test results would positively correlate with higher amplitudes of both mentioned N and P ERP waveforms. More specifically, results were predicted to follow the ERP patterns of in-group and outgroup (or

congruent and incongruent) stimuli and the ERP waveforms discussed in previous research. Paired with the IAT results, this would also indicate that participants have a faster latency for familiar or in-group stimuli and a slower latency for unfamiliar or out-group stimuli. Thus, higher D scores were predicted to correlate with more biased answers and thus more drastic amplitudes. Familiar or congruent stimuli in this research were Black faces paired with weapons or white faces paired with harmless objects; and unfamiliar or incongruent stimuli were white faces paired with weapons and Black faces paired with harmless objects. To test these hypotheses participants completed a Weapons IAT test while wearing the EEG skullcap.

Method

Participants

Thirty-four participants (*M* age=20 years of age) recruited through the Connecticut College SONA system or through advertising of the study in specific phycology classes on campus. Of these participants 8 identified as male, 21 identified as female and zero identified as gender non-binary or queer. Of these participants 17 identified as white, 7 as Black, 1 as Hispanic and 4 as Asian. Five participants chose not to fill out their demographics forms. Four participants did not complete the EEG trial and thus their trials were discarded. Four other participants were excluded from analysis due to noisy EEG data In total there were 26 participants whose EEG data was analyzed. Each participant was compensated either with class credit or payment.

Research Design

The current study employed a within subjects, experimental design with some correlational aspects, in which participants completed the Weapons IAT while wearing an EEG skullcap. Behavioral data (i.e. participants' response times and error rates for the IAT) were recorded as they completed the task and later were calculated into a D score which will be

explained further in the IAT section below. While participants did the IAT task an EEG system recorded their data. Each participant received sections of the IAT in a randomized order, however the test blocks always remained the same. All participants signed informed consent forms and received debriefing forms after the experiment was over.

Implicit Association Test

Participants completed a modified version of the Weapons Implicit Association Test (see Appendix A). This test comes from Project Implicit, a laboratory based at Harvard University ("Project Implicit", 2019). The purpose of this particular IAT was to measure the strength of association between the race of the face in the image and the type of object presented. Participants completed a version of the test that was very similar to that which is offered through Harvard's online database titled "the Weapons IAT" with added trials for each block (permission to use the stimuli from the weapons IAT was obtained from Project Implicit). The stimuli were reprogrammed so that the IAT could be implemented using EEG software. IAT stimuli were viewed on a standard computer monitor. After the participant's completion of the IAT the software generated a D score was calculated for each participant by averaging latencies during each block. D score calculations will be explained in more detail below. There were 7 blocks total in the task.

Directions were provided to each participant before the task began. The task was explained step by step, with the directions first prompting the participant to press the letter "I" and then the next screen prompting them to press the letter "E". These are the two letters that were used during the task that corresponded to different categories of stimuli. After the initial directions were presented block 1 began. At the beginning of the first block, auditory feedback (i.e. a beeping noise) was provided when participants answered incorrectly. If the answer was

correct there was no beep. This was to ensure that any behavioral data collected was as accurate as possible, and that there were no incorrect trials due to the participants lack of understanding of the task. At the beginning of each block the computer would very clearly indicate which category went with which key. For example, for block one the computer screen would indicate "for white faces (press I) and for Black Faces press (E), press the space bar to begin" as seen in *Appendix A*.

For the first block stimuli was solely comprised of faces. The second block contained on harmless objects like apples and pencils, and weapons like guns and hand grenades. Participants were asked to differentiate between the two. The directions read "Press I for weapons and E for harmless objects. Block three was the first block to contain both types of categories, the directions read "White faces and weapons, Press I, Black Faces and harmless objects, press E." Block three was the first incongruent block, meaning that white faces and weapons are not stereotypically together and Black faces and harmless objects (thought to represent harmlessness) are not stereotypically thought of together. Examples of harmless objects would be apples, pens, erasers, and examples of weapons would be things like guns and knives, as seen in *Appendix B*. Block number four was another incongruent block and the directions for this block read, "for white faces and weapons, press I, Black faces and harmless objects, press E." Block five simply switched keys used for each group. The directions read "For Black faces press I, for white faces press E "and to be sure that participants understood the change that had occurred the directions also read "watch out the labels have changed". Block six proceeded in a similar fashion to block three and read "for Black faces and weapons, press I, for white faces and harmless objects, press E". Block six is the first example of a congruent block, due to the pairing of Black faces and weapons. The congruent nature of this block lies in the relationship between Black faces and weapons and white faces and harmless objects or harmlessness. Black faces and

violence or weapons are congruent in nature because they are so often ingrained into our social norms and stereotypes (even if the phenomenon of Black violence is not a complete representation of reality) (Lundberg et al., 2018). The last block, block seven was another example of a congruent block and the directions read the same as block number six. A full presentation of the described block pattern and corresponding keys can be found in *Appendix A*. The order blocks 3 or 4 and 6 or 7 were counterbalanced meaning that some participants had blocks 3 and 4 as congruent while other had those blocks as incongruent. The same thing occurred for blocks 6 and 7. The basic order for the blocks can be seen in *Appendix A*.

After the initial directions disappeared from the first screen before each block began the screen would still read which key corresponded with which category in the upper corners of the screen for the remainder of each block. For example, in block one in the upper left hand corner of the screen the computer would read "E for Black faces" and in the upper right hand corner the screen would read "I for White faces". These prompts would stay throughout the block. In the meantime, photos that belonged to one of the four categories mentioned above, depending on the block, would appear in the center of the screen and would remain there until a response was given. The photos were relatively small, about two inches by two inches and some varied slightly in with or length. Each photo was exactly the same from the set of stimuli given by Project Implicit ("Project Implicit", 2019). However, due to the fact that the current study has an EEG component more photos were repeated in order to create the number of trials necessary. Examples of these images can be found in *Appendix B*.

In order to score the IAT accurately blocks are broken down into practice and test blocks. Blocks 3, 4 and 6, 7 were test blocks, more test blocks were added in this study similarly to other EEG and IAT studies (Schiller et al., 2016). Test blocks contained 200 trials and were

the only blocks counted in calculating the final D score. The amount of trials in the test blocks were much greater than in a normal IAT test without an EEG component. This is because EEG can be a very noisy measure and thus the more trials that can be analyzed, the more reliable the results. Blocks 1, 2, and 5 were practice blocks. These blocks contained only 20 trials and were not included in the final D score. The purpose of the practice blocks was to allow the participant to rehearse using specific keys in response to the stimuli on the screen, to be positive that they understood the task and to practice answering as fast as possible. The block order was always randomized for counter balancing purposes.

Calculating the D-Score

A D score is calculated by determining the difference in mean reaction time between trials from pooled incongruent and pooled congruent blocks. The D score is typically reported as the behavioral measure of association between two categories. IAT D scores are computed as the mean difference divided by the pooled standard deviation between test blocks (Westfall, 2015). The most and simple equation used to compute D scores is as follows D= (M2-M1)/SD. D score is only calculated for correct trials and incorrect trials are corrected by using a value of +600ms (so that the value was above the mean). (FreeIAT: How It Works.,2018, May 8). M1 and M2 are representative of the two categories, in the current study M2 represents incongruent blocks while M1 represents congruent blocks. Therefore, a higher or more positive D score indicates more bias. The idea behind this is that it is more difficult for the participant to respond to incongruent trials, and thus the latency in these trials should be longer, resulting in a higher final D score (Nosek et al. 2016).

After the data were gathered, the trials with a latency greater than 10000 milliseconds were removed, and the accuracy of each participant was calculated. An overall D score was also calculated for each participant. Based on the final D scores, participants were divided into two groups using a median split. Specifically, those 13 participants whose D scores were above the median were assigned to the "high implicit bias group" and those 13 participants who fell below the median were designated as the "low implicit bias group. IAT measures have been reported to have greater internal consistency than other implicit tests. For example, Cronbach's α values typically range from .7 to .9 (Nosek et al. 2016). The excellent internal consistency is the reason for using the median split to separate participants into high and low D score groups and remain confident in the statistical analysis to follow.

EEG/ERP

Prior to the administration of the IAT, an EEG cap was placed and fitted to each participant. Electrical brain activity was recorded from electrodes placed on the participant's scalp using a 32 channel ActiCHamp system provided by Brain Products. These electrodes are designed to measure summed electrical activity from the brain. Electrode gel was inserted with a blunt needle around the electrodes in the EEG cap. Electrodes were moved and secured in order to reduce impedance. EEG recording was not carried out in an electrically shielded environment. However, the ActiCHamp system creates the same effect by using active electrodes. A ground electrode was used. TP10, the left mastoid, was used as an online reference. The average of TP9 (right mastoid) and TP10 (left mastoid) were used as reference offline after all the data was collected. The reference sights were chosen based on previous ERP literature to allow comparability of results to other studies (Healy, Boran & Smeaton, 2015).

EEG clean up and ERP averages were calculated by using EEG lab and ERP lab in order to look at specific components and for clean-up (A Delorme & S Makeig, (2004); Lopez-Calderon et al., (2014). After the data was collected it was referenced to the average of the mastoids. A bandpass filter was used to filter out frequencies above 30hz and below .1 hz. Artifact removal filtered out movements which exceeded 70 mV or contained other noise-like artifacts. Artifacts such as eye blinks and facial muscle were rejected. Individual electrodes that were too noisy were interpolated based on data from their closest neighbors.

Baselining and Epoching

A base line of -200ms to 0ms pre-stimulus was used for the stimulus-locked epoch extraction. Baselining removes noise like inter-subject differences and slow-drifts. This allows the inter-subject measures to be comparable as ERP amplitudes relative to zero, which is the baseline (Healy, Boran & Smeaton, 2015). In this particular study and other studies that use IAT and EEG together it is important to understand that the IAT experimental design may give rise to some confounds. Due to the participants understanding that the IAT measures bias it is important to select the baseline correctly (Healy, Boran & Smeaton, 2015). This is why in previous literature and the current study the baseline was selected for -200-0 ms. The data were then epoched and averaged for each event be categorized for congruent and incongruent blocks.

ERP Time-windows and Channel Selection

Topographic variations of ERP activity in the IAT literature implicate a number of fundamental ERP components. Regions of interest were chosen based on previous EEG and IAT literature. Electrodes F3, FZ, F4, FC3, FCZ, FC4, CZ, C3, and C4 were of most interest in each ERP waveform. ERP waveforms were collected from -200ms to 796 ms however, eye blinks

were removed and ERP averages were closely examined 0-400ms. Time windows for each waveform were selected. For N400, 350-450ms was selected (Wang et al., 2010; Healy, Boran & Smeaton, 2015). For N100 110-150ms (Healy, Boran & Smeaton, 2015). For P200 which was relevant to the current study 160-230ms was the selected time window (Volpert, 2012). For N200, which was also relevant to the current study, 250-310 ms and for P300 330-450 ms (Healy, Boran & Smeaton, 2015). In the current study negative ERP was plotted downwards.

Results

Behavioral results

A single sample t-test was conducted to determine if the D score was statistically significant greater than zero, which would be indicative of a pattern of implicit bias on the IAT. Overall, D scores (which indicates latency and bias) were marginally significantly greater than zero (M=.119, SD= .309), $t(25)$= 1.923, p=.06. A higher D score would mean that the participant had a harder time (greater latency) answering in incongruent (White faces and weapons) blocks then in a congruent (Black faces and weapons) blocks. The results of this t-test show marginal significance and indicate that our sample did show the expected bias. These results support the original hypothesis and the previous literature that participants would show bias on the IAT(Greenwald, Benaji & Nosek, 2003).

Choosing significant ERPs

After review of all of the raw data and scalp maps using MATLAB (MathWorks, Inc., 1996). P200 and N200 waveforms seemed of most significance for examination. These two ERP waveforms were predicted to be of most importance, however other amplitudes that were predicted to show large differences between groups did not show the expected differences and thus were not analyzed further. Scalp mapping of electrodes shows some visual differences

between amplitudes in congruent and incongruent groups for P200 and N200 for electrodes F3, FZ, F4, FC3, FCZ, FC4, CZ, C3, and C4, this can be seen in figure1and figure 2. In figure 1, not much of a difference was observed between congruent and incongruent groups. However, this scalp map allowed for understanding of which electrodes might be most useful to analyze more closely by showing the range of amplitudes in different regions on the map. This way the regions that were most active for this component were visible so precise electrodes could be chosen. In figure 2, the largest difference between groups occurred with the amplitude difference between viewing Black faces or white faces. A smaller difference between white and Black faces occurs in the congruent (Black faces and weapons) grouping relative to the incongruent (white faces and weapons) grouping. These visual results support the original hypothesis that incongruent groups should show different results then congruent groups. Similar to previous studies, medial electrodes were identified as playing a large role with the current data (Healy, Boran & Smeaton, 2015).

Specific ERPS were also chosen based on figures comparing all waveforms in specific scenarios. Overall congruent and overall incongruent waveforms from channels F3, FZ, F4, FC3, FCZ, FC4, CZ, C3, and C4 were analyzed in order to pinpoint specific components that showed significant differences between the two groups. This can be visualized in figure 3 where both overall congruent and incongruent groups are shown. The biggest visual differences occurred at P200 and N200 and thus these two components were of most relevance to the current study. The difference between the incongruent data and congruent data were then examined to see if differences in wave forms existed between D score levels (high D score, more biased, and low D score, less biased). This is shown in Figure 4 as the difference between block types (incongruent subtracted from congruent) in each separate group (high and low D score). Again, it is clear the

biggest visual differences occurred at the P200 mark and end during N200 time windows and that there are noticeable visual differences between groups on the graph. The ERPs were chosen based on visual inspection of components that had already been focused on in the literature and pinpointed in the original hypothesis. This is important to note due to the issue of multiple comparisons (Healy, Boran & Smeaton, 2015).

The current study also examined the neurological differences when participants saw a Black face on the screen or a white face on the screen in order to understand whether ERP differences stem from actual bias or just from visual differences. As predicted in the original hypothesis, and as displayed in figure 5, it is clear that the color of the stimuli does drastically affect the ERP amplitude especially in the P200 and N200 time windows. This could be due to the Hillyard Principle or due to underlying bias and will be discussed further below (Stevens & Bavelier, 2012). Each statistical analysis was conducted on either overall data or one of two wave forms, P200 and N200.

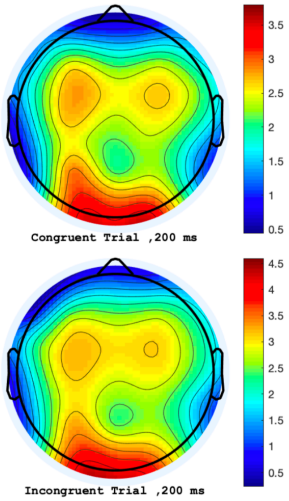

Figure 1. P200 map of the electrodes and comparison between congruent and incongruent groups averaged across all participants. Groups are similar at 200ms.

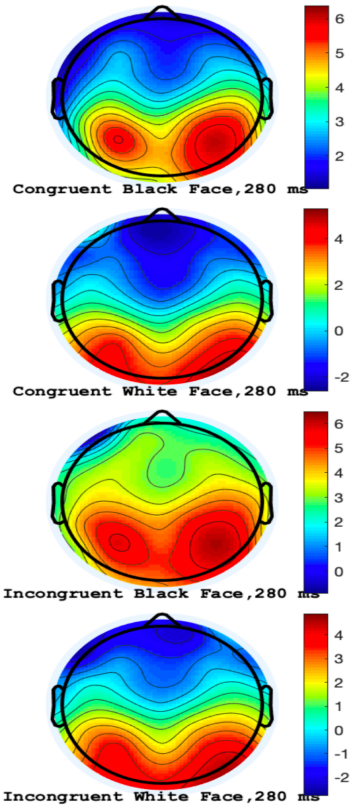

Figure 2. N200 map of electrodes on the scalp. White faces in comparison to Black faces in both congruent and incongruent groups at 200 ms.

P200 results

A 2 (block type) x 2 (D score grouping) factorial ANOVA was conducted to compare the effect of block type on the P200 ERP peak amplitude (which was the dependent variable) in separate (high and low) D score groupings. Block type was treated as a within subjects variable and included two levels - congruent and incongruent. D score groups were treated as a between subjects variable and contained two levels high median (high D score, more biased) and low median (low D score, less bias). There was no significant main effect for block type alone F (1, 24) =.011, p=.916. As expected, there was also no main effect of D score grouping F (1, 24) =.037 p=.85. This supports the hypothesis that neither variable would have an effect alone but when put together an effect may occur.

An interaction effect, between block type and D score grouping was significant, F (1, 24) =12.27, p= .002. These results support the original hypothesis and indicate that congruent and incongruent blocks did have a different P200 waveform depending on high or low median D score groupings. The difference between D score grouping in the congruent block category is smaller than in the incongruent block category but small visual differences can be seen at P200 and N200 (figure 3). For the incongruent block there is a visual difference between less and more biased D score groups around 200ms (figure 4). Surprisingly, the difference between congruent and incongruent waveforms was higher for low bias or low D score groups which can be explained in the correlations section below. The same idea can also be visualized in the overall difference between congruent and incongruent (overall incongruent peak amplitude minus overall congruent peak amplitude) compared with D score for all participants (the negative correlation graph; see figure 6).

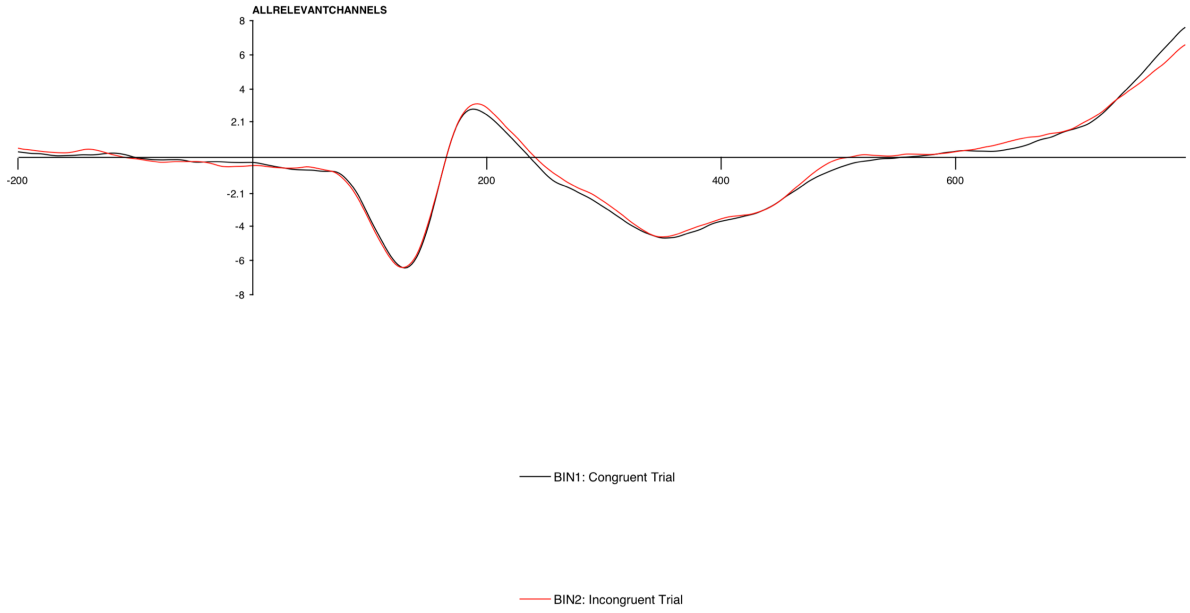

Figure 3. Overall congruent and Overall incongruent.

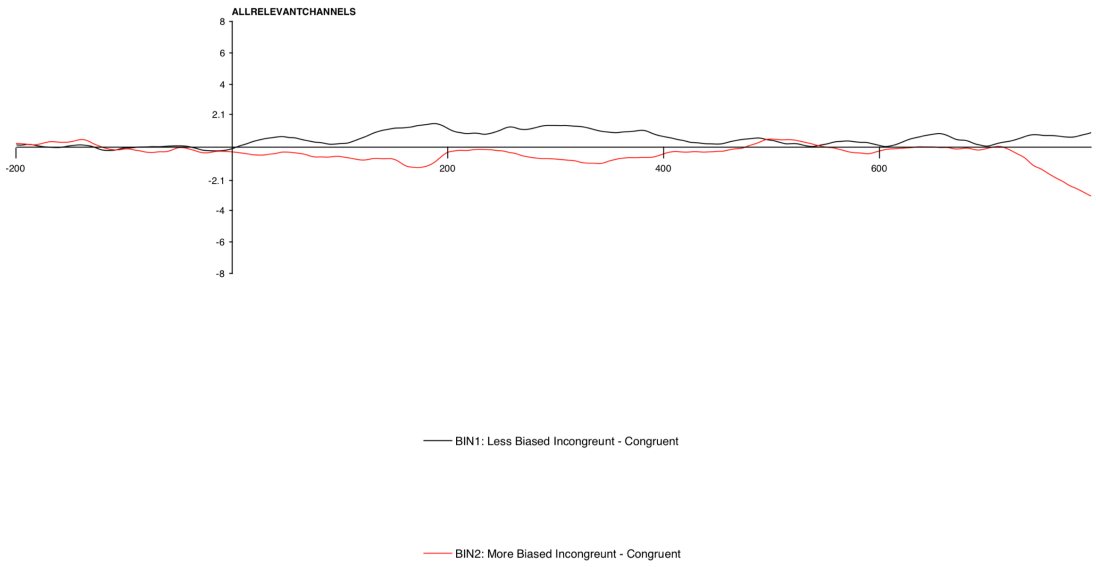

Figure 4. Incongruent subtracted from congruent scores. These differences were averaged in high median and low median D score groups.

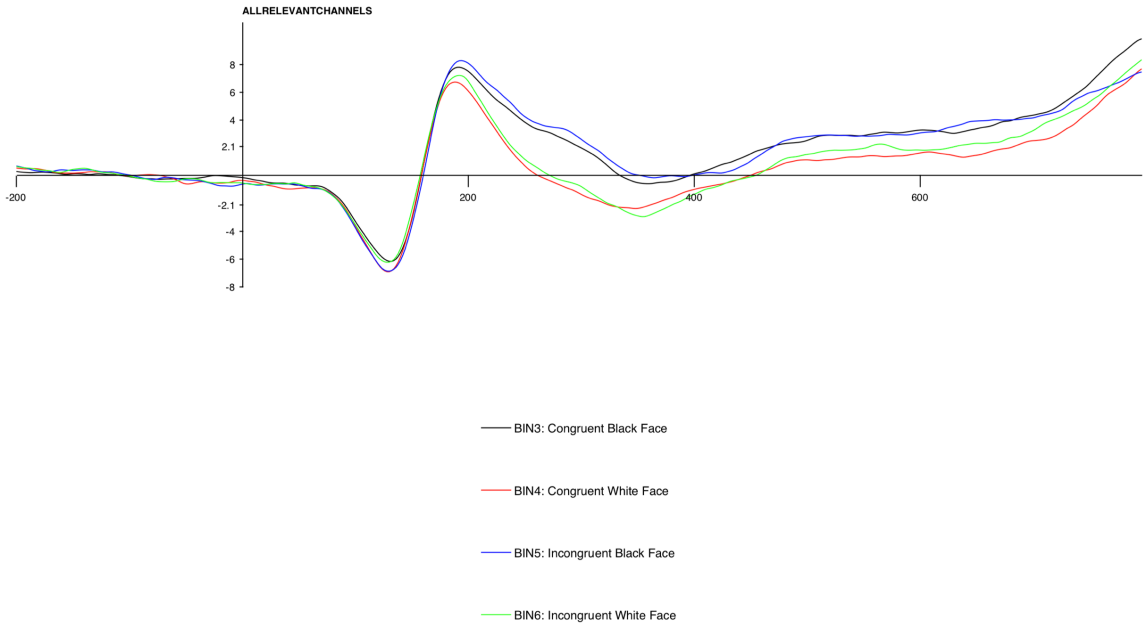

Figure 5. Shows the difference between Black faces and white faces separately in the congruent group versus the incongruent group.

N200 Waveform results

A 2 (block type) x 2 (race) x 2 (D score grouping) factorial ANOVA was conducted to compare the effect of block type and race on the N200 ERP mean amplitude (the dependent variable) in high median and low median D score groups. Block type was treated as a within subjects variable and included two levels- congruent and incongruent. Race was also a within subjects variable and included two levels- Black faces and white faces. D score grouping was treated as a between subjects variable and contained two levels high median (high D score, more biased) and low median (low D score, less bias). There was no significant effect for block type F (1, 24) = 1.189, p=.286 or for interaction between block type and D score grouping F (1, 24) =.008, p=.928. There was also no significant effect for face type and D score grouping F (1, 24) =.1.77, p=.195, although this was the interaction effect that was closest to being significant. This result was not supportive of the original hypothesis because N200 mean amplitude did not change significantly when different types of faces (Black or white) were shown in either high median and low median D score groups. Lastly, no interaction effects were found for block type and face type F (1, 24) =.036, p=.852 and for the interaction effect of block type, face type and D score grouping F (1, 24) =.481, p=.495. No interactions were significant. The lack of interactions effects also did not support the hypotheses.

There was a significant main effect of the face type (Black or white) on N200 mean amplitudes. F (1, 24) = 37.622, p<.05. These results supported the original hypothesis and indicated that the race of the face (Black or white) did significantly affect the N200 amplitude. This may be because of actual underlying bias or, because of the Hillyard principle which will be explained further in the discussion section (Stevens & Bavelier, 2012). The same analysis was

not done for P200 due to visual inspection of the data that suggested that N200 mean amplitude was the best component for comparison.

Correlational Analysis: The Relationship Between ERP data and Behavioral Data

Correlational analyses were performed in order to understand the relationships between the primary dependent variables. In order to facilitate these analyses, we subtracted the average of the peak amplitude of incongruent trials from the of the peak amplitude of congruent trials for P200 for each individual participant and calculating the difference. We then correlated this score with overall D scores. The results indicated that D scores were significantly negatively correlated with the difference in overall congruent and incongruent (incongruent peak amplitude minus congruent peak amplitude) trials $r = -.519$, $p < .05$ (figure 6). These results indicated that the D score had a significant relationship with the difference between individuals congruent and incongruent amplitudes. The correlation was such that the higher the individual D score the smaller difference between incongruent and congruent their peak amplitudes were (figure 6). Meaning that the more biased a person seems to be, according to the behavioral measures, the smaller the difference between block type amplitudes. This result is the exact opposite of the originally proposed hypothesis that more biased participants would show greater changes between block type amplitudes (congruent and incongruent). This may be due to a conscious or subconscious attempt to exert cognitive control over implicit bias. P200 and N200 have both been shown to be affected by cognitive control and will be addressed in more detail in the discussion (Kumar et al., 2010; Correll et al., 2006). While the results did not support the original hypothesis, they should still be taken into consideration due to the fact that they were statistically significant.

The results also indicated that there was no statistically significant correlation between D scores and overall congruent block amplitudes for P200, $r = .033$,$p = .913$ (figure 7). There was also no statistically significant correlation between D scores and overall incongruent blocks for P200 as seen in figure 7, $r = -.178$, $p = .386$, (figure 8). These results show that individual block type did not significantly correlate with D scores but that the difference between block type amplitudes (the difference between overall congruent peak amplitudes and overall incongruent peak amplitudes) did negatively correlate with D scores. This did not support the hypothesis. However, it may make sense that individual block type did not correlate with D score due to the nature of the analyses which clearly shows that it is the relationship between block types that are more important rather than individual block type. D scores were expected to correlate positively with difference in block type peak amplitudes. Instead a negative correlation was observed (figure 6).

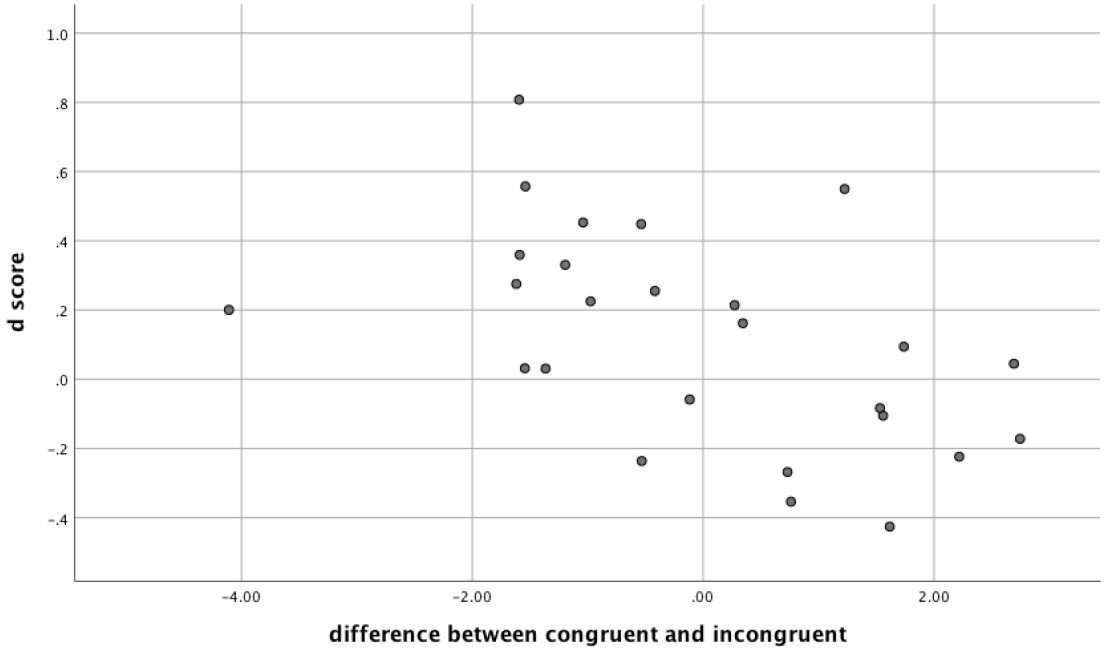

Figure 6. Shows a significant negative correlation between the difference between overall congruent score and overall incongruent score versus the D score for all participants.

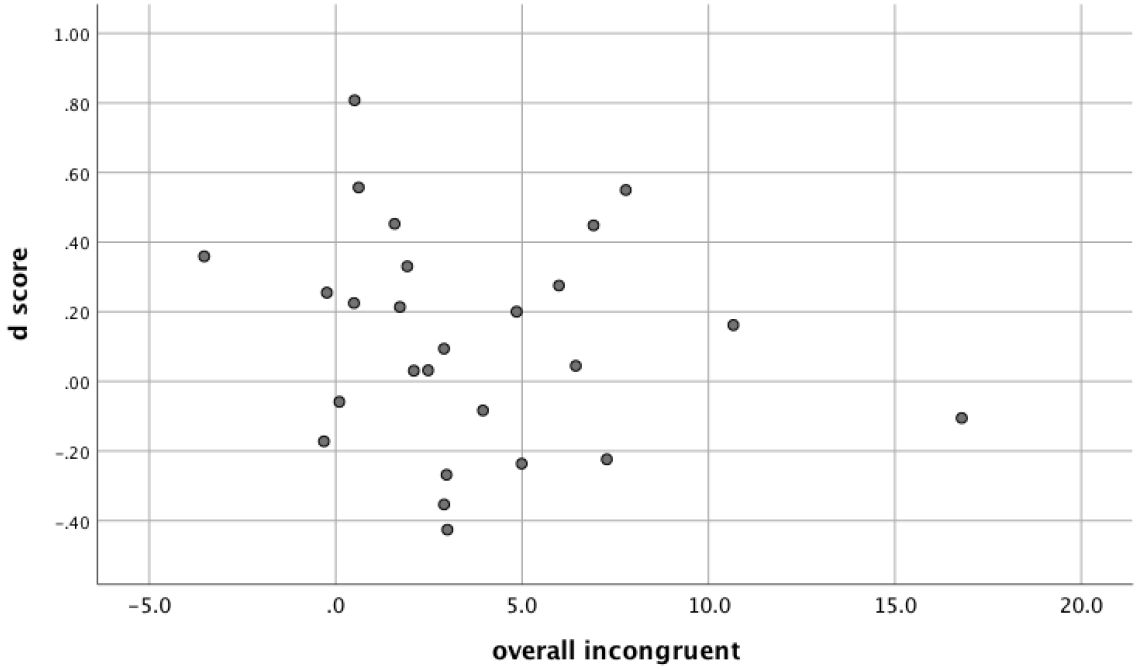

Figure 7. Shows no correlation between overall congruent scores and D score for all participants.

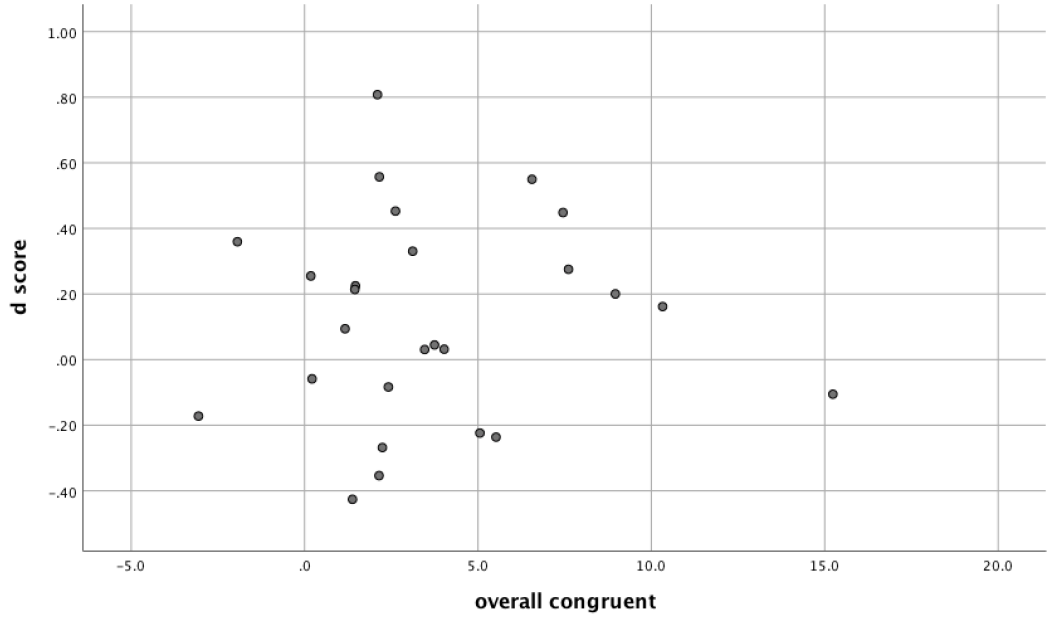

Figure 8. Shows no correlation between overall congruent scores and D score for all participants.

Discussion

The current study examined the relationship between weapons bias (calculated in the form of an IAT D score) and ERP amplitudes. Behavioral data replicated previous findings that D scores were marginally significantly greater than zero, thereby demonstrating a pattern of implicit bias on the IAT. These results indicated that latencies were longer for incongruent groups, and shorter for congruent groups, meaning that seeing white faces with weapons (incongruent) was more likely to spark a biased response (shown in behavioral and ERP data), relative to seeing Black faces with weapons (congruent). This was likely due to the stereotypes often seen throughout society that have become familiar to us. The incongruent group functioned as the unfamiliar group (white faces and weapons) whereas the congruent group functioned as the familiar or stereotyped group (Black faces and weapons). These results not only support previous IAT findings but also indicated that racial weapons bias is very prevalent. The results also showed that the population used in the current study did display the type and degree of bias that had been originally hypothesized to exist in the broader population (Greenwald, Benaji & Nosek, 2003; Correl et al., 2006).

P200 and N200 were selected as the most relevant ERP waveforms to analyze after scalp maps and ERP raw data were analyzed. As expected, these two waveforms were the most visually different between congruent and incongruent groups and thus were most indicative of bias (figure 4). These results are similar to those of many previous studies that isolated P200 and N200 as waveforms to show the most significant bias (Healy, Boran & Smeaton, 2015; Correl et al., 2006). One study in particular conducted by Correl et al. (2006) yielded very similar results. After conducting an experiment where participants were asked to shoot targets in a video game

that were white, Black, armed and unarmed researchers concluded the N200 and P200 appeared most significantly affected by race and bias.

As hypothesized there was an interaction effect between block type (congruent or incongruent) and D score grouping (high and low D score) for the P200 component. The specificity of what the interaction effect would look like was not hypothesized however, and P200 peak amplitude was larger on incongruent trials for only the less bias D score group. This suggests that D score, or level of implicit bias, likely influenced how a participant viewed congruent and incongruent blocks (see figure 3). Also, this result suggests that this may have occurred on a less significant level in the N200 component, but future studies would need to be conducted to confirm this. In the future, these findings could serve as a way in which researchers could examine ERP data in order to predict the level of bias (high or low) of a participant. ERP research could also be vital to helping to change biased habits and thought processes by monitoring them via EEG.

Sensitivity to race in the form of different faces in the N200 waveform in the current study may have been indicative of racial bias. As hypothesized changes in ERP amplitudes would occur when different types of faces (black or white) were viewed. However, the change in amplitude at the N200 waveform for different races occurred for all participants regardless of high or low D score, which did not support the original hypothesis that more biased participants would have a greater change in amplitude between races. Although not supported by the original hypothesis a study conducted by He et al (2009), which was already discussed in the introduction, does show a similar ERP pattern when participants looked at stimuli that contained different Black and white faces. Why did the measure behavioral bias (IAT) not completely match with what was anticipated for the N200 ERP waveform? One reason that would partially

explain both the current study and the study conducted by He et al (2009) is that the Hillyard principle may be at play. The Hillyard principle assumes that if the differences in the N200 amplitude cannot be attributed to a specific cognitive process (like bias), then perhaps the N200 differences could be attributed sensory differences between Black and white faces, alongside the bias (Hillyard et al., 1973). The change in ERP amplitudes to different stimuli was first demonstrated by Hillyard who used ERP to examine auditory attention. He found that different tones elicited different amplitudes in many ERP waveforms (Stevens & Bavelier, 2012). Hillyard found that stimuli effect selective attention and neural processing differently depending on what the stimuli are. The original landmark study found that the most drastic differences in response to a set of varied pitches were at the N100 component (negatively peaking between 80 and 120 ms) (Hillyard et al., 1973; Stevens & Bavelier, 2012). Other ERP studies found that the same phenomenon occurred most frequently at the P200 level and also occasionally at the N200 level (Stevens & Bavelier, 2012). In the case of the current study both underlying bias and sensory changes may have contributed to the same effect. Although not predicted in the original hypothesis, it is important to consider why the differences in amplitudes for Black and white faces are occurring at the N200 level. These results may very well be due to the underlying racial bias that has been taken into account during the behavioral analysis of the IAT. These results could also indicate a neurological change that was simply elicited by the difference in color of the faces shown during the task that could be categorized under the Hillyard principle. Another likely possibility is that these results are due to both underlying bias and the Hillyard principle.

Although there was a main effect for race on N200 mean amplitudes likely due to the both the Hillyard principle and underlying implicit bias, there was no interaction effect for D score grouping and face type (race; Black or white). Although there was no significant

interaction between D score and race the interaction that did occur between race and amplitude was the most significant out of the three interaction effects calculated for N200 mean amplitudes. This information, paired with the significant main effect for race, may possibly indicate that there is an underlying bias attributing the changes in mean amplitude when white or Black faces are shown. However, though this bias may exist, the limitations of the current study, such as the population size or population demographics of the current study inhibited that bias from being brought to light. Due to the fact that race and amplitude had the effect that was closest to significant, but was still not very significant, another explanation could be that the observes overall reaction to face type regardless of behavioral bias is the same due to cognitive control. In this scenario cognitive control would be exerted by the more biased group to overcome the underlying bias (Kumar et al., 2010). This may be a valid interpretation because previous research has made it clear that cognitive control is reflected in both the N200 and P200 amplitudes (Kumar et al., 2010). Either explanation would require future research to pinpoint the cognitive processes that may be underlying these neurological results.

The correlational analyses indicated that there was no correlation between overall peak congruent amplitude and D score, or between overall peak incongruent amplitude and D scores for P200. D scores were not expected to significantly correlate with block types. However, D scores and the difference between overall peak incongruent and overall peak congruent trials for P200 (incongruent peak amplitude minus congruent peak amplitude) were expected to be positively correlated. In other words, it was predicted that the bigger the difference between the two block types the larger the D score would be. However, the opposite result was observed, showing a significant negative correlation between D score and difference in block type for P200. This significant negative correlation is important in that it shows the relationship between

D score and difference in block type. However, it was the opposite of the original hypothesis indicating that the higher (more biased) the D score the smaller the difference between block type.

Other IAT, EEG studies have found significant correlations between D score and individual block type for N200 and less so for P200 (Healy, Boran & Smeaton, 2015). For example, a study conducted by Healy et al. (2015) also hypothesized that D scores would correlate with respective incongruent blocks due to proactive cognitive control and other top-down control processes. Their results indicated that N200 amplitudes showed a negative correlational relationship with D score that highlighted the possible involvement of cognitive control in the IAT (Healy, Boran & Smeaton, 2015). Other studies also point to both N200 and less frequently P200 as evidence of cognitive control during a task (Kumar et al., 2010). The effect of cognitive control, when completing the IAT, would greatly change the internal goals of each participant. The conscious or even subconscious attempt to suppress one's own bias would result in very different ERP amplitudes and patterns. For example, if cognitive control did elicit the negative rather than positive correlation between D score and the difference between groups, then the results would indicate two levels of cognitive function, one being the implicit bias and one being the attempt at controlling the bias either consciously or subconsciously when being asked to perform on a weapons IAT. In fact, one major criticism of the IAT itself is that it could be less valid due to the fact that if participants know their implicit bias is being examined they will try to control the results and show less bias (Oswald et al., 2013). This could explain the findings in the current study.

One other possibility for seeing a negative correlation between D score and difference between block types is the effect of education. Perhaps being educated about the nature of bias

may have some effect on neurological responses to those biases. It is possible that less biased participants (low D score group) show less biased answers on the IAT because throughout their education they have learned to process and diminish their implicit bias. However, the education around bias could also create different neurological responses for the two separate groups (congruent and incongruent) to be seen more separately and more clearly but without being biased towards one of the groups. This would indicate that people in the less biased group are possibly more aware of race, and thus are drawing on their learned skills to minimize their bias, but in order to do that their brains more clearly recognize the difference in race in each stimuli. The neurological difference seen in the current study could be due to the awareness required to minimize the bias.

It is also possible that having a more biased score makes some the participants see each group similarly, even though there is a clear difference between the congruent and incongruent groups or Black or white faces. This would also indicate that having less implicit bias allows the participant to see each group for what it truly is. For example, if the lower D score people were truly less biased because the saw less of a difference between groups (congruent or incongruent) then we would expect that their responses Black and white faces would be the same. However, we see just the opposite in Figure 5, where it is clear that lower D score participants see an even bigger difference in Black and white faces. This makes it possible that the low D score group is showing less behavioral bias but greater ERP changes between congruent and incongruent groups, due to the fact that they have less implicit bias. Being aware of their implicit bias, possibly through education, could create this effect.

To further explore these ideas, it would be important in future research to look at the effects of education on implicit bias. Maybe becoming less biased requires explicit thinking

about your implicit biases more rather than trying to ignore differences between groups. This could be tested in a future experiment that begins with 30 people who have already been flagged as high D score participants. The experiment could then be set up so that participants need to go through some number of days of explicit training about their implicit bias and then their IAT and ERP data could be compared before and after. An experiment like this could determine if in fact a larger difference in amplitude between the two groups (congruent and incongruent) was not only tied with cognitive control but also with a new ability to see individual groups as individuals rather than trying to see them as the same in order to try to diminish their bias.

Lastly, it is important to recognize that while the results obtained in the current study were the opposite of the proposed hypotheses, the expected bias was indeed observed on the IAT and did have some relationship to the EEG measure. Future research should examine exactly how and for what reason this occurred. Furthermore, these results could imply that almost all people suffer from some form of bias, but some people may have greater cognitive control or have learned to change their perceptions of their own bias via education or experience. If cognitive control can effectively reduce bias (not just in IAT results but also in practice) it may be interesting in the future to teach cognitive control rather than just trying to completely erase the underlying bias. It is possible that beginning with cognitive control, people may find their underlying bias diminishing as well over time. This could potentially push white society to put in the work required to recognize their own biases and begin to change them.

Limitations

There were several limitations to the current study. One limitation was the way in which the IAT itself can serve as both a measure with high validity, and yet and can also be criticized

for the eliciting cognitive control (Oswald et al., 2013). Cognitive control could be a confounding variable, however, because it might obscure the relationships between IAT scores and EEG; more research would need to be done in order to see the real role of cognitive control in IAT and EEG.

It is also possible that cognitive control may change the outcome of the second measure that is paired with the IAT, in this case that would be EEG (Oswald et al., 2013). Another limitation of the study had to do with the number of participants and the demographics of the participants at Connecticut college. While originally there were 34 participants in the study, after retaining only those participants' who completed all measures, only 26 viable data sets remained. This is a number that is on the lower side for EEG studies (Healy, Boran & Smeaton, 2015). It is also important to recognize that while there was some racial diversity and gender diversity in the population used in the study, Connecticut College is a predominantly white institution and the sample studied in the present research also reflects that demographic trend. This would allow for the findings to be generalized to similar demographic settings, however the findings of the present study could probably not be generalized to a broader and more diverse population. In order to arrive at broader conclusions, studies should aim to recruit more people of color. Moreover, EEG as an imaging technique, in the field of studying bias, is also limited due to the novelty of this type of research and the fact that bias and racism is most often examined using fMRI and other imaging techniques that can see where in the brain bias may originate, rather than the electrical frequencies that arise from it (Stanley et al., 2008)

Future Research

In the future it would be important to conduct more research employing a broader variety of different samples in order to more fully understand the role that implicit bas may play in EGG imaging. Additional research examining cognitive control of implicit bias could be conducted in order to understand which wave forms are affected by control of bias rather than cognitive control more generally. This could be done be telling participants about their bias before the EEG. A future study could contain three groups, one control group and two experimental groups. The first experimental group could simply be made aware of their bias and then asked to complete the IAT/ EEG study. In the second group participants would be made aware of their bias and researchers could exaggerate the negative effects of their bias maybe even tell participants that their bias is extremely detrimental to their relationships and how they go about their lives. This way research could understand at which levels cognitive control takes place and changes EEG outcomes. This is similar to a study mentioned in the introduction conducted by Correll et al. (2006) where participants completed a video game task and some groups were made aware of their bias and others were not. This format for research would allow for a clearer distinction between cognitive control of bias and actual underlying bias.

Studies could also employ different versions of the IAT. For example, some that studies might examine different types of implicit bias and determine whether there is a more general waveform pattern for bias or in-group/ out-group related studies. The same idea could be used to examine if cognitive control may appear in other forms of bias that are not weapons related as well. Lastly, it would be important to eventually conduct a meta-analysis of all the different types of brain imaging research, in order to determine the strength of the relationship between

IAT behavioral data and fMRI and EEG. This way researchers could begin to see even more broad patterns of implicit or explicit bias in the brain.

Practical and Theoretical Implications

The current study is important in that it contributes to a growing field of EEG and IAT research. One important outcome of this study is that it serves as another reminder that weapons bias, and in a more general sense implicit bias, is commonplace. Now, not only can bias begin to be seen in ERP patterns, but we can begin to understand how bias and cognitive control interact with one another in a way that might be analyzed in EEG imaging. The more the science community understands implicit bias and how it occurs in the brain, the closer the world will be to educating people and making them aware of their bias, in order to hopefully one day gain greater control over it. This control would entail not only the disempowering of subconscious bias, but perhaps result in a greater understanding of the different forms of racism that continue to persist.

One inspiration for this study was the famous "Blue-Eyes, Brown-Eyes" experiment conducted by Jane Elliot in her grade school class room in 1968 and further discussed in the PBS film *A Class Divided* (2013). In this experiment Elliot divided her grade school class in half; those with brown eyes and those with blue eyes. On the first day of the experiment she told the class that the children with blue eyes were smarter, nicer, neater, and overall better children then those with brown eyes. On the first day she gave the children with blue eyes more praise and more privileges then those with brown eyes. Children with brown eyes had to wear specific collars around their necks and were criticized often that day. On the second day the roles were completely reversed so that brown eyes children were getting the same praise and rewards as the

blue eyes children had the day before and the blue eyed children were being scolded and scapegoated. Both days had similar results; children who were made out to be inferior took on the image and behaviors of actually inferior students. For example, they performed poorly on tests and other school work. The "superior" students became rude and seemed to enjoy discriminating against the "inferior" group (Erickson, 2004).

This in-class experiment is fascinating on its own, but it informed research that followed decades later in film (Erickson, 2004). The broadcast of PBS's *A Class Divided* presented Jane Elliot's work to a national audience in the United States. This program also showed the profound impact her demonstration had on the children in her class who were involved in the study. The PBS video is often cited as a longitudinal study for Elliot's work and shows participants in the original experiment talking about how they feel that they lead very different lives as a result of being a part of the "Blue-Eyes Brown-Eyes" experiment. Participants, as adults, have discussed having a deeper understanding of how biases and in-group/out-group dynamics are detrimental to the people around us and how easy it is to fall into the trap of not seeing ones' own biases. These adults, who took part in the film *A Class Divided*, also reported being in inter-racial marriages, living in more diverse neighborhoods and even having lives dedicated to social justice.

Understanding the ways in which in-group/out-group biases can be taught and understood may hopefully change how human beings learn about each other in school and relate in other contexts as well. This would have many implications for both international and domestic politics, the practice of medicine, the understanding of history, and almost any other conceivable human endeavor. Hopefully, in the not so distant future, understanding how in-group/ outgroup bias manifests itself at the neurological level could help control the human proclivity to treat each other unjustly, and perhaps even save lives.

References

Amodio, D. M., & Ratner, K. G. (2011). Mechanisms for the Regulation of Intergroup Responses: A Social Neuroscience Analysis. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780195342161.013.0048

A Delorme & S Makeig (2004) EEGLAB: an open source toolbox for analysis of single trial EEG dynamics (pdf, 0.7 MB) Journal of Neuroscience Methods 134:9-21 *Includes details of EEGLAB ICA and time/frequency methods.*

Artyom Zinchenko, Philipp Kanske, Christian Obermeier, Erich Schröger, Sonja A. Kotz, Emotion and goal-directed behavior: ERP evidence on cognitive and emotional conflict, *Social Cognitive and Affective Neuroscience*, Volume 10, Issue 11, November 2015, Pages 1577–1587, https://doi.org/10.1093/scan/nsv050

Brown, R. (2010). *Prejudice: its social psychology*. Malden (MA): Wiley-Blackwell.

Billig, M. and Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1), pp.27-52.

Caharel, S., Montalan, B., Fromager, E., Bernard, C., Lalonde, R., & Mohamed, R. (2011). Other-race and inversion effects during the structural encoding stage of face processing in a race categorization task: An event-related brain potential study. *International Journal of Psychophysiology,79*(2), 266-271. doi:10.1016/j.ijpsycho.2010.10.018

Campanella, S., Quinet, P., Bruyer, R., Crommelinck, M., & Guerit, J. (2002). Categorical Perception of Happiness and Fear Facial Expressions: An ERP Study. *Journal of Cognitive Neuroscience,14*(2), 210-227. doi:10.1162/089892902317236858

Campanella, S., Gaspard, C., Debatisse, D., Bruyer, R., Crommelinck, M., & Guerit, J.-M.

(2002). Discrimination of emotional facial expressions in a visual oddball task: An ERP study. *Biological Psychology, 59*(3), 171-186.doi: 10.1016/S0301-0511(02)00005-4

Chen, A., Xu, P., Wang, Q., Luo, Y., Yuan, J., Yao, D., & Li, H. (2008). The timing of cognitive control in partially incongruent categorization. *Human Brain Mapping*, *29*(9), 1028–1039. doi: 10.1002/hbm.20449

Chekroud, A. M., Everett, J. A., Bridge, H., & Hewstone, M. (2014). A review of neuroimaging studies of race-related prejudice: Does amygdala response reflect threat? *Frontiers in Human Neuroscience,8*. doi:10.3389/fnhum.2014.00179

Cikara, M., Farnsworth, R. A., Harris, L. T., & Fiske, S. T. (2010). On the wrong side of the trolley track: Neural correlates of relative social valuation. *Social Cognitive and Affective Neuroscience,5*(4), 404-413. doi:10.1093/scan/nsq011

Correll, J., Urland, G. and Ito, T. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology*, 42(1), pp.120-128.

Derks, B. (2013). *Neuroscience of prejudice and intergroup relations*. New York, NY: Psychology Press.

Elliott, J., Yale University., WGBH (Television station : Boston, Mass.), & PBS DVD (Firm). (2003). *A class divided*. New Haven, Conn.: Yale University Films.

Elsner, B., Jeschonek, S. and Pauen, S. (2013). Event-related potentials for 7-month-olds' processing of animals and furniture items. *Developmental Cognitive Neuroscience*, 3, pp.53-60.

Erickson, I. M. (2004). Fighting Fire with Fire: Jane Elliott's Antiracist Pedagogy. *Jstore*, 1–14.

Retrieved from https://www.jstor.org/stable/pdf/42978385.pdf?refreqid=excelsior:300ca4dfc77a504468540aa28a82a803

Fatal Force: 2019 police shootings database. (2018, January 2). Retrieved from https://www.washingtonpost.com/graphics/2019/national/police-shootings-2019/.

FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, *18*(1), 19. doi:10.1186/s12910-017-0179-8

FreeIAT: How It Works. (2018, May 8). Retrieved from https://meade.wordpress.ncsu.edu/freeiat-home/freeiat-how-it-works/.

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *Journal of General Internal Medicine,22*(9), 1231-1238. doi:10.1007/s11606-007-0258-5

Greenwald, A. G., & Nosek, B. A. (2016). Health of the Implicit Association Test at Age 3. doi:10.31234/osf.io/sv8bv

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. doi:10.1037/0022-3514.85.2.197

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the Lowest of the Low. *Psychological Science,17*(10), 847-853. doi:10.1111/j.1467-9280.2006.01793.x

He, Y., Johnson, M. K., Dovidio, J. F., & Mccarthy, G. (2009). The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. *Social Neuroscience,4*(5), 426-442. doi:10.1080/17470910902949184

Healy, G., Boran, L. and Smeaton, A. (2015). Neural Patterns of the Implicit Association Test. *Frontiers in Human Neuroscience*, 9.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical Signs of Selective Attention in the Human Brain. *Science*, *182*(4108), 177–180. doi:10.1126/science.182.4108.177

Hobson, H. M., & Bishop, D. V. (2017). The interpretation of mu suppression as an index of mirror neuron activity: Past, present and future. *Royal Society Open Science,4*(3), 160662. doi:10.1098/rsos.160662

Ibanez, A., Baker, P., & Moy, A. (2012). Event-Related Potential Studies of Cognitive and Social Neuroscience. *Neuroimaging - Cognitive and Clinical Neuroscience*. doi:10.5772/24514

Ibáñez, A., Gleichgerrcht, E., Hurtado, E., González, R., Haye, A., and Manes, F. F. (2010). Early neural markers of implicit attitudes: N170 modulated by intergroup and evaluative contexts in IAT. *Front. Hum. Neurosci.* 4:188. doi: 10.3389/fnhum.2010.00188

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in Cognitive Sciences,13*(12), 524-531. doi:10.1016/j.tics.2009.10.002

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports,6*(1). doi:10.1038/srep39589

Katsumi, Y., & Dolcos, S. (2018). Neural Correlates of Racial Ingroup Bias in Observing Computer-Animated Social Encounters. *Frontiers in Human Neuroscience,11*. doi:10.3389/fnhum.2017.00632

Korn, H. A., Johnson, M. A., & Chun, M. M. (2012). Neurolaw: Differential brain activity for

Black and White faces predicts damage awards in hypothetical employment discrimination cases. *Social Neuroscience,7*(4), 398-409. doi:10.1080/17470919.2011.631739

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature, Neuroscience*, *15*(7), 10.1038/nn.3136. http://doi.org/10.1038/nn.3136

Kumar, N., Sood, S., Singh, M., Beena, & Sakshi (2010). Effect of acute moderate exercise on cognitive event-related potentials n100, p200, n200, and interpeak latencies. *Indian journal of psychological medicine*, *32*(2), 131–135. doi:10.4103/0253-7176.78511

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in human neuroscience, 8, 213

Luck, S. J. (2005). *An Introduction to Event-Related Potentials and their Neural Origins (Chapter 1)*. Cambridge: MIT Press.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Lundberg, G., Neel, R., Lassetter, B. and Todd, A. (2018). Racial bias in implicit danger associations generalizes to older male targets. *PLOS ONE*, 13(6), p.e0197398.

MathWorks, Inc. (1996). MATLAB : the language of technical computing : computation, visualization, programming : installation guide for UNIX version 5. Natwick :Math Works Inc.,

McIntosh, P. (2003). White privilege: Unpacking the invisible knapsack. In S. Plous (Ed.), *Understanding prejudice and discrimination* (pp. 191-196). New York, NY, US: McGraw-Hill.

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience &*

Biobehavioral Reviews,37(8), 1530-1536. doi:10.1016/j.neubiorev.2013.06.002

Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The Neural Dynamics of Facial Identity Processing: Insights from EEG-Based Pattern Analysis and Image Reconstruction. *Eneuro*. doi:10.1523/eneuro.0358-17.2018

Nunspeet, F. V., Ellemers, N., Derks, B., & Nieuwenhuis, S. (2012). Moral concerns increase attention and response monitoring during IAT performance: ERP evidence. *Social Cognitive and Affective Neuroscience,9*(2), 141-149. doi:10.1093/scan/nss118

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

Payne, B. K. (2006). Weapon Bias. *Current Directions in Psychological Science*, *15*(6), 287–291. doi: 10.1111/j.1467-8721.2006.00454.x

Payne, B.K., Lambert, A.J., & Jacoby, L.L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in racebased misperceptions of weapons. *Journal of Experimental Social Psychology,* 38, 384–396.

Polich J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, *118*(10), 2128–2148. doi:10.1016/j.clinph.2007.04.019

ProjectImplicit. (n.d.). Retrieved from https://implicit.harvard.edu/implicit/contact.html

Puente, A. E. (1992). Assessment of Possible Neural Substrates. *Schizophrenic Disorders,*161-177. doi:10.1007/978-1-4757-2159-1_8

Peck, T. C., Seinfeld, S., Aglioti, S. M., & Slater, M. (2013). Putting yourself in the skin of a

black avatar reduces implicit racial bias. *Consciousness and Cognition,22*(3), 779-787. doi:10.1016/j.concog.2013.04.016

Schiller, B., Gianotti, L. R. R., Baumgartner, T., Nash, K., Koenig, T., & Knoch, D. (2016). Clocking the social mind by identifying mental processes in the IAT with electrical neuroimaging. *Proceedings of the National Academy of Sciences*, *113*(10), 2786–2791. doi: 10.1073/pnas.1515828113

Stanley, D., Phelps, E., & Banaji, M. (2008). The Neural Basis of Implicit Attitudes. *Current Directions in Psychological Science,17*(2), 164-170. doi:10.1111/j.1467-8721.2008.00568.x

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, *2*. doi: 10.1016/j.dcn.2011.11.001

Sheng, F., Liu, Y., Zhou, B., Zhou, W. and Han, S. (2013). Oxytocin modulates the racial bias in neural responses to others' suffering. *Biological Psychology*, 92(2), pp.380-386.

Sur, S., & Sinha, V. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal,18*(1), 70. doi:10.4103/0972-6748.57865

Tajfel,H. and Wilkes, A. (1963). Classification and Quantitative Judgment. *British Journal of Psychology*, 54(2), pp.101-114.

The Stanford Open Policing Project. (n.d.). Retrieved from https://openpolicing.stanford.edu/findings/.

ThoughtCo. (2019). *What Does Implicit Bias Really Mean?*. [online] Available at: https://www.thoughtco.com/understanding-implicit-bias-4165634 [Accessed 23 Sep. 2019].

Valla, L. G., Bossi, F., Calì, R., Fox, V., Ali, S. I., & Rivolta, D. (2018). Not Only Whites: Racial Priming Effect for Black Faces in Black People. *Basic and Applied Social Psychology*, *40*(4), 195–200. doi: 10.1080/01973533.2018.1462185

Volpert, H. (2012). Unique Neural Correlates of Implicit and Explicit Bias. (1), pp.1-37.

Wang, Y., Zhang, Z., Bai, L., Lin, C., Osinsky, R., & Hewig, J. (2017). Ingroup/outgroup membership modulates fairness consideration: Neural signatures from ERPs and EEG oscillations. *Scientific Reports,7*, 39827. doi:10.1038/srep39827

Westfall, J. (2015, May 31). Retrieved from http://jakewestfall.org/misc/D_bound.html.

Xu, F. K., Nosek, B. A., Greenwald, A. G., & Lofaro, N. (2014, February 28). Experiment Materials. Retrieved from https://osf.io/k6g3m/

Appendix A

Weapons Implicit Association Test, Block Break down

Directions: what letter to press for what category and to go as fast as possible

Block	Content
1	White faces (press I) Black Faces press (E) -
2	Weapons (press I) Harmless Objects press (E)
3 test	White faces and weapons (Press I) Black Faces and harmless objects (press E)
4 test	White faces and weapons (press I) Black faces and harmless objects (press E)
5	Black Faces (press I) white faces (press E) “watch out the labels have changed”
6 test	Black faces and weapons (press I) white faces and harmless objects (press E)
7 test	Black faces and weapons (press I) white faces and harmless objects (press E)

Appendix B

Examples of each category; Black and white faces, Weapons and Harmless Objects