THE YALE LAW JOURNAL

ELIZABETH INGRISELLI

Mitigating Jurors' Racial Biases: The Effects of Content and Timing of Jury Instructions

ABSTRACT. This Note examines, through an experimental design, whether juror biases against black defendants are explained by aversive racism theory or social identity theory and whether procedural justice can be used to decrease biases. The Note also examines whether the timing of debiasing jury instructions affects judgments of guilt. The experiment finds that pre-evidence instructions result in lower judgments of guilt than post-evidence instructions. In addition, aversive racism theory, but not social identity theory or procedural justice, explains guilt judgments. The experiment has implications for both the content and timing of jury instructions in trials.

AUTHOR. Yale Law School, J.D. expected 2016; Princeton University, A.B. 2011. I am especially thankful to Tom Tyler for his help with my research design, to Justin Sevier for his help with data analysis, and to the students in Yale Law School's Empirical Research Seminar for their feedback on my research design and the interpretation of my results. I am also grateful to Nicole Shelton, Susan Fiske, and John Darley for introducing me to the prejudice and discrimination literature in social psychology. I would additionally like to thank William Nguyen, as well as Devon Porter, Meng Jia Yang, and Rachel Bayefsky for their helpful and insightful suggestions throughout the editorial process.

NOTE CONTENTS

INTRODUCTION

The Sixth Amendment of the United States Constitution guarantees the accused the "right to a . . . trial, by an impartial jury . . . ."[1] This assumes that jurors can divorce themselves from any biases that they might have and decide cases based on relevant evidence. Such assumptions are crucial to the legitimacy of the criminal justice system. If jurors do not carefully examine evidence in order to make the best decision possible, then defendants' constitutional right to an *impartial* jury will be undermined. Recent research into juror decision making shows that jurors have difficulty remaining impartial and often exhibit racial biases. Although this research has shown that black defendants are more likely to be found guilty than white defendants for the same crime,[2] no research thus far has investigated the underlying mechanism that causes these racial biases. Instead, many researchers have attributed such effects, post hoc, to aversive racism theory.[3] However, social identity theory can also plausibly explain such results. Briefly, aversive racism theory is based on the idea that today racism is more implicit and unconscious than explicit and conscious. Many individuals are not explicitly or consciously racist, and do not wish to be so, but have implicit, unconscious biases that emerge when they are unable to monitor their biases.[4] Social identity theory, on the other hand, identifies a more conscious process, whereby individuals exhibit biases favoring ingroup members and disfavoring outgroup members in order to raise the status of their ingroup.[5]

This Note builds on prior research and attempts to pinpoint which of these two theories – aversive racism theory or social identity theory – better explains

---

1. U.S. CONST. amend. VI.

2. *E.g.*, Ellen S. Cohn et al., *Reducing White Juror Bias: The Role of Race Salience and Racial Attitudes*, 39 J. APPLIED SOC. PSYCHOL. 1953, 1964 (2009); *see also* Samuel R. Sommers & Phoebe C. Ellsworth, *Race in the Courtroom: Perceptions of Guilt and Dispositional Attributions*, 26 PERSONALITY & SOC. PSYCHOL. BULL. 1367, 1378 (2000) [hereinafter Sommers & Ellsworth, *Race in the Courtroom*] (finding that white jurors are more likely to express racial biases when race is not salient); Samuel R. Sommers & Phoebe C. Ellsworth, *White Juror Bias: An Investigation of Prejudice Against Black Defendants in the American Courtroom*, 7 PSYCHOL. PUB. POL'Y & L. 201, 225 (2001) [hereinafter Sommers & Ellsworth, *White Juror Bias*] (same); Denis Chimaeze E. Ugwuegbu, *Racial and Evidential Factors in Juror Attribution of Legal Responsibility*, 15 J. EXPERIMENTAL SOC. PSYCHOL. 133, 140 (1979) (noting that white jurors rate black defendants as more culpable than white defendants when evidence is ambiguous).

3. *See* Sommers & Ellsworth, *White Juror Bias*, *supra* note 2, at 225; *see also* Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1376 (stating that aversive racism theory may explain the experimental result that white jurors have racial biases).

4. *See* discussion *infra* Part I.A.

5. *See* discussion *infra* Part I.B.

juror biases. To determine which theory better explains biases and can simultaneously combat biases, the experiment in this Note manipulated the content of jury instructions designed to reduce bias (referred to below as "debiasing jury instructions") and measured mock jurors' judgments of the guilt of black defendants. This experiment also tested whether instructions based on procedural justice can decrease juror biases. Procedural justice posits that individuals are more likely to comply with legal rules, such as jury instructions, if they view legal decision-making processes as fair.[6] In addition to testing whether the content of jury instructions can mitigate biases, this experiment examines whether high or low explicit race salience[7] and whether the presentation of instructions before or after the presentation of evidence affect jurors' guilt judgments of black defendants.

Part I provides a theoretical overview of aversive racism theory, social identity theory, and procedural justice. Part II then provides empirical evidence in support of the race-based aversive racism theory and social identity theory. It investigates three shortcomings in aversive racism theory that the present experiment attempts to resolve. It also proposes that the results from prior experiments are compatible with social identity theory and urges that, as a result, this theory should be investigated as a potential cause of juror biases.

Part III lays out the experimental hypotheses derived from aversive racism theory, social identity theory, procedural justice, and the timing of jury instructions. In short, these hypotheses are as follows. First, if aversive racism theory explains biases, then high explicit race salience and jury instructions that remind jurors of their egalitarian views should decrease judgments of guilt for aversive racists only. Second, if social identity theory explains biases, then low explicit race salience and jury instructions that attempt to increase participants' self-esteem should mitigate judgments of guilt for all participants. Third, if procedural justice reduces juror biases, then debiasing jury instructions emphasizing procedural justice should decrease judgments of guilt for all participants. Finally, the timing hypothesis predicts that the debiasing instructions, if they work, should be more effective at decreasing biases and guilt judgments when presented pre-evidence than when they are presented post-evidence.

Part IV presents the experimental methods used, including the sample of participants, experimental design, materials, and procedure. Part V presents the results of the experiment. The results suggest that aversive racism theory, but not social identity theory, explains jurors' racial biases, and that instruc-

6. *See* discussion *infra* Part I.C.

7. In the experiment, explicit race salience refers to the experimental condition in which race was specifically discussed and there was a picture of the defendant that made his race more apparent. *See* discussion *infra* Part IV.B.

tions tailored to aversive racism theory are likely to reduce such biases. Additionally, the results do not support the hypothesis that jury instructions focused on procedural justice reduce biases. Finally, the results largely support the timing hypothesis: jurors were less biased when the debiasing instructions were presented before the evidence. Part VI discusses the results, limitations of the experiment, and directions for future research. The Note concludes by outlining the practical implications of the experimental results and suggests that judges ought to include debiasing elements based on aversive racism theory in their jury instructions and present these instructions before the evidence phase of the trial.

I. THEORETICAL ACCOUNTS OF JURORS' RACIAL BIASES

A. Aversive Racism Theory

Aversive racism is a modern form of racism in which whites exhibit implicit biases – biases of which they are unaware but that have discriminatory effects – against blacks.[8] Although researchers have recognized that aversive racism theory is not limited to whites and blacks,[9] the theory has typically focused on whites' biases against blacks.[10] Aversive racists are those who "regard themselves as nonprejudiced and nondiscriminatory; but, [they] almost unavoidably[] possess negative feelings and beliefs about blacks."[11] They are high in implicit racism, yet low in explicit racism – in other words, they are biased but are

8. Samuel L. Gaertner & John F. Dovidio, *The Aversive Form of Racism*, *in* PREJUDICE, DISCRIMINATION, AND RACISM 61, 62 (John F. Dovidio & Samuel L. Gaertner eds., 1986).

9. Samuel L. Gaertner & John F. Dovidio, *Understanding and Addressing Contemporary Racism: From Aversive Racism to the Common Ingroup Identity Model*, 61 J. SOC. ISSUES 615, 619 (2005).

10. This Note focuses on white juror biases against black defendants, rather than biases between other racial groups, for two interrelated reasons. First, scholars have framed aversive racism theory as explaining whites' biases against blacks rather than biases between other groups. *See generally* JOEL KOVEL, WHITE RACISM: A PSYCHOHISTORY, at xi, xxiii-xxiv (1970) (coining the term "aversive racism" and using it to explain whites' biases against blacks). *See also* Gaertner & Dovidio, *supra* note 9, at 617-19 (describing aversive racism in terms of whites' biases against blacks, without mentioning other racial groups). Second, because of this framing, the empirical research on juror biases has focused on white biases against black defendants rather than between other groups. Recall that most researchers have attributed white juror biases to aversive racism theory, *see supra* note 3 and accompanying text, so there has been little exploration of whether white jurors may have biases against other racial groups as well, even though social identity theory would predict such biases. I have therefore focused on these racial categories in order to respond and contribute to this body of literature.

11. Gaertner & Dovidio, *supra* note 8, at 62.

unaware of their biases. Because they are low in explicit racism and thus wish to be non-racist in accordance with their egalitarian views, aversive racists suppress their prejudice toward blacks when they are made aware of their biases,[12] which occurs when race is made salient.[13] When race is not salient, and "norms are ambiguous or conflicting, discrimination is often exhibited"[14] due to whites' implicit biases. Aversive racism theory can be applied to the legal realm: if race is made salient in the courtroom, then white jurors often suppress their negative attitudes toward black defendants in an attempt to appear egalitarian, thereby reducing racially biased decision making.[15] Conversely, in trials in which racial factors are not salient, white jurors' implicit biases against black defendants will emerge.[16]

B. Social Identity Theory

According to social identity theory, individuals categorize others into ingroups or outgroups and favor members of the ingroup to enhance their own self-image.[17] Social identity theory is derived from three assumptions about individual behavior and society. First, individuals wish to have high self-esteem, which is defined as a positive self-concept.[18] Second, social groups have positive and negative qualities, and individuals' personal social identities are derived in part from society's evaluations of their social groups.[19] Third, the evaluation of an individual's ingroup is determined by social comparisons to other groups.[20] These three assumptions lead to the core theoretical principles of social identity theory: (1) individuals wish to achieve a positive social identity through (2) comparisons that favor the individual's ingroup and disfavor an outgroup, and (3) if an individual's ingroup lags behind an outgroup, the individual will leave the ingroup or try to improve the ingroup's image.[21]

---

12. *Id.*

13. Samuel R. Sommers & Phoebe C. Ellsworth, *"Race Salience" in Juror Decision-Making: Misconceptions, Clarifications, and Unanswered Questions*, 27 BEHAV. SCI. & L. 599, 601 (2009).

14. Gaertner & Dovidio, *supra* note 8, at 85.

15. *See, e.g.*, Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1371.

16. *See, e.g., id.*

17. *See* Henri Tajfel & John Turner, *An Integrative Theory of Intergroup Conflict*, *in* THE SOCIAL PSYCHOLOGY OF INTERGROUP RELATIONS 33 (William G. Austin & Stephen Worchel eds., 1979).

18. *Id.* at 40.

19. *Id.*

20. *Id.*

21. *Id.*

At the core of social identity theory is the idea that individuals wish to have a positive image of themselves and the status of one's ingroup contributes to this image.[22] Thus, they favor the ingroup over an outgroup to improve the image of the ingroup, which in turn improves their own self-image.[23] The logic of social identity theory predicts that individuals may similarly favor their ingroup in criminal trials in order to increase their self-esteem.[24] If ingroup members are suspected of having committed a crime, individual members of the group may be more lenient in judging the ingroup member. Otherwise, an ingroup member's conviction might suggest the proclivity of the ingroup toward crime. This in turn would lower the status of the ingroup and harm the individual's self-esteem, because the status of the ingroup affects judgments of one's self-esteem. Thus, social identity theory can plausibly explain why white jurors judge black defendants to be guiltier than white defendants. White jurors are harsher toward black defendants because those defendants belong to an outgroup; by acting harshly toward an outgroup, white jurors indirectly raise the status of their ingroup.

C. *Procedural Justice*

Unlike aversive racism theory and social identity theory, procedural justice is a race-neutral theory that focuses on impartiality and fairness and thus does not *explain* racial biases. Nevertheless, it may be employed to *combat* jurors' racial biases. According to this theory, individuals view the justice system as legitimate if the process by which it reaches outcomes is perceived to be fair.[25] This legitimacy then leads to compliance with legal rules.[26] Therefore, procedural justice may lower racial biases by the following mechanism: if jurors are

22. *See id.*

23. *See id.*

24. This logic assumes that evidence is ambiguous. This assumption is reasonable in this context, as many cases that actually go to trial (as opposed to being resolved by a guilty plea) involve ambiguous evidence. When there is strong evidence incriminating a member of one's ingroup, then individuals are actually *more* likely to convict the ingroup member, as they are motivated to separate a clearly guilty member from the ingroup to preserve the ingroup's status. *See infra* text accompanying notes 74-78.

25. *See generally* JOHN THIBAUT & LAURENS WALKER, PROCEDURAL JUSTICE: A PSYCHOLOGICAL ANALYSIS (1975). Legitimacy is defined as a "sense of obligation or willingness to obey authorities." Margaret Levi et al., *Conceptualizing Legitimacy, Measuring Legitimating Beliefs*, 53 AM. BEHAV. SCIENTIST 354, 356 (2009); *see also* Jason Sunshine & Tom R. Tyler, *The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing*, 37 LAW & SOC'Y REV. 513, 514 (2003) ("Legitimacy is a property of an authority or institution that leads people to feel that that authority or institution is entitled to be deferred to and obeyed.").

26. Levi et al., *supra* note 25, at 356.

primed with the idea that the process by which they reach a verdict is fair, they will see juror decision making as more legitimate, and therefore will be more likely to comply with jury instructions. So if jury instructions emphasize the fairness of the decision-making process and stress impartiality or lack of bias, jurors will be more likely to follow the instructions and suppress any biases that they might have.

II. EMPIRICAL EVIDENCE FOR THE THEORIES

A. Aversive Racism Theory

This Part and the next suggest that the race and juror decision-making literature has three significant limitations, which drive this Note's experimental design. First, the literature's conceptualization of explicit race salience does not explain studies that have found no bias when race is not salient—when jurors' biases should emerge. Second, prior studies introduced the confounding factor of an interracial crime when making race salient. Finally, and most significantly, prior studies did not measure participants' racism levels. This Part discusses each limitation, and how the experimental design attempts to address it, in turn.

Aversive racism theory centers on the hypothesis that white jurors will be nonbiased when race is made salient but will exhibit biases when race is not salient. In order to understand why aversive racism theory explains a lack of bias when race is salient, one must first understand what race salience means. Most studies have conceptualized race salience to refer to a situation in which "the nature of the trial emphasizes race as a central issue."[27] This conceptualization ignores more subtle instances of race salience, and this Part addresses this gap by extending the previous definition. This extension is important for two reasons: (1) it suggests that prior studies that ostensibly contradict aversive racism theory may actually support the theory; and (2) it provides the impetus for the debiasing jury instructions in one of the conditions in this Note, as explained below.

The new conceptualization is as follows: for race to be salient, it need not be a "central issue" in a trial. Subtle reminders of race can impact juror decision making: under aversive racism theory, all that matters is that jurors are reminded of their egalitarian views before they deliberate. By this new definition, race salience need not be confined to a central, explicit issue mentioned in the courtroom; if racial issues are salient in the societal context in which a trial takes place, this implicit salience should also decrease biases. In fact, there need

27. Sommers & Ellsworth, *supra* note 13, at 608.

not even be implicit race salience as long as jurors are reminded of their egalitarianism. As a result, this Note posits that egalitarianism – rather than explicit or implicit race salience – is the central explanatory factor in aversive racism theory. In other words, race need not be salient to decrease jurors' biases: race salience, whether implicit or explicit, decreases biases because it reminds aversive racists of their egalitarian views. Because egalitarianism, rather than race salience, is the causal factor, triggering egalitarianism directly would be simpler and more powerful. This Part will first examine empirical evidence in support of the traditional definition of race salience, which I call "explicit race salience." Then it will examine evidence in favor of the broadened definition of race salience, which I call "implicit race salience," and I will explain how this new definition provides the impetus for the jury instructions in this study.

1. *Traditional Definition: Explicit Race Salience*

According to Samuel Sommers and Phoebe Ellsworth, race salience refers to the idea that race is the "central issue" of a trial.[28] In line with this definition, Sommers and Ellsworth manipulated race salience in one of their studies in the following manner: (1) in the "race salience" condition, racially charged language was used during the crime or race had motivated the crime itself; and (2) in the "no race salience" condition, no racially charged language was used nor had race motivated the crime.[29] Sommers and Ellsworth found that white jurors' guilt judgments of black and white defendants did not differ when race was salient.[30] When race was not salient, however, white jurors convicted the black defendant more often than the white defendant.[31] Subsequent studies have replicated this race-salience effect.[32] These results support aversive racism theory; explicit race salience reminds jurors of their egalitarian values, causing them to suppress their negative feelings toward blacks, and this in turn leads to equal treatment of black and white defendants. When race is not explicitly salient, however, white jurors' implicit biases emerge, leading to disparate treatment of white and black defendants. Although many studies have found white juror bias when race is not explicitly salient, other studies have failed to repli-

28. *Id.*

29. Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1369, 1373.

30. *Id.* at 1369, 1373-74.

31. *Id.* at 1374.

32. Cohn et al., *supra* note 2, at 1961; Sommers & Ellsworth, *White Juror Bias*, *supra* note 2, at 217.

cate this effect.[33] At first glance, this inconsistency suggests that aversive racism theory is limited in its scope. However, I will argue that the studies that have not found white juror bias when race is not explicitly salient have actually created implicit race salience. Because implicit salience serves the same function as explicit race salience in reminding jurors of their egalitarian views, the lack of white juror bias in those studies may actually support aversive racism theory.

2. *New Definition: Implicit Race Salience*

I posit that prior literature has discussed implicit race salience in two forms, though the literature has not framed these discussions in terms of implicit salience: (1) racial tension in society that is implicitly brought into the courtroom; and (2) awareness of the race of one's fellow jurors.

a. *Race Salience in a Societal Context*

A study conducted in Los Angeles during the O.J. Simpson trial by Paul Skolnick and Jerry Shaw provides evidence for the first type of implicit race salience: racial tension in society that is implicitly brought into the courtroom. The researchers presented jurors with a case in which a man was accused of murdering his ex-wife.[34] They found that white jurors were equally likely to convict white and black defendants.[35] Because race was not made explicitly salient during the trial (the facts of the case insinuated that the man murdered his wife out of emotional rage with no mention of race or racial motivations),[36] these findings, when taken at face value, seem to contradict the aversive racism theory framework. Although the facts did not mention race, the researchers claim that racial tensions during the study were high because it was conducted in Los Angeles during Simpson's civil trial, and the facts in the study were very similar to Simpson's case.[37]

Race salience during pre-trial publicity may lead to white jurors' suppression of racial biases.[38] In Skolnick and Shaw's experiment, white jurors may

33. *E.g.*, Paul Skolnick & Jerry I. Shaw, *The O.J. Simpson Criminal Trial Verdict: Racism or Status Shield?*, 53 J. SOC. ISSUES 503, 513 (1997); *see also* Samuel R. Sommers, *Race and the Decision Making of Juries*, 12 LEGAL & CRIM. PSYCHOL. 171, 172 (2007).

34. Skolnick & Shaw, *supra* note 33, at 506.

35. *Id.* at 511.

36. *Id.* at 506-08.

37. *Id.* at 514-15.

38. *See* Sommers & Ellsworth, *supra* note 13, at 606 (stating that more research needs to be performed to determine the impact of race salience from pre-trial publicity).

have suppressed their biases because the racially salient social context in Los Angeles reminded them of their egalitarian views. This link can potentially explain why whites sometimes act in an nonbiased manner even when race is not made explicitly salient at a trial: if race is implicitly salient through the societal context in which the trial takes place, then jurors may act as if race had been made salient at the trial itself, as either form of salience reminds them of their egalitarian views.

b. Implicit Race Salience in Racially Diverse Juries

Implicit race salience may also explain why diverse juries become less biased in their decision making: the race of the minority jurors makes race implicitly salient for white jurors, reminding them of their egalitarianism and thereby decreasing their biases. In support of this idea, a study found that majority-white juries were more likely to convict a Hispanic defendant than a white one, whereas majority-Hispanic juries were equally likely to convict a Hispanic defendant and a white one.[39] These findings support the new definition of implicit race salience because juries with Hispanic majorities make race salient for whites, even if race is not an explicit central issue at trial, leading them to suppress any biases that they may have. In sum, because race is more salient in majority-Hispanic juries, whites, in accordance with aversive racism, will act in a less biased manner than in cases in which race is less salient, such as in majority-white juries. In this respect, the effects of race salience fall along a continuum: in majority-white juries there are still minorities on the jury, but not enough to make race as salient as when Hispanics are the majority, leading to discrepant outcomes between the two juries.

An alternative explanation is that white jurors' biases may not have changed; the discrepant outcomes between majority-Hispanic and majority-white juries may instead be caused by a shift in power on the jury. One could argue that the majority-Hispanic jury was nonbiased because those in power were nonbiased, and thus their opinions were discussed more often. Likewise, on the majority-white jury, there were more advocates for a biased opinion, leading to a more polarized and biased decision.

Although this power hypothesis is a plausible explanation, group polarization based on power structure alone cannot account for the different verdicts of majority-Hispanic and majority-white juries. Pre-deliberation attitudes among whites become less biased when they expect to sit on racially heterogeneous ju-

39. Dolores Perez et al., *Ethnicity of Defendants and Jurors as Influences on Jury Decisions*, 23 J. APPLIED SOC. PSYCHOL. 1229, 1256 (1993).

ries,[40] and this suppressed bias then leads whites to discuss different issues than when they are on homogeneous juries.[41] Samuel Sommers compared the pre-deliberation attitudes, deliberation content, and verdicts of homogenous all-white juries and heterogeneous white-and-black juries.[42] He found that when whites were told that they would serve on a heterogeneous jury, they were significantly less likely to think that the black defendant was guilty than were white jurors who were told they would serve on a homogeneous all-white jury.[43] In addition, researchers have found that whites anticipating serving on racially diverse juries are more attuned to race-related concepts; they complete word-stem completion tasks with more race-related words than do those anticipating serving on homogeneous juries.[44] The expectation of serving on a racially diverse jury may therefore act as a form of implicit race salience that reminds whites of their egalitarian views and inhibits their negative feelings toward blacks. This implicit salience then leads whites to discuss different issues than when they serve on homogeneous juries.[45]

Anticipating serving on racially diverse juries also leads to more in-depth processing of race-relevant information and the suppression of biases. White jurors expecting to interact in racially diverse groups perform better on reading comprehension tests that contain race-relevant questions than whites expecting to interact in all-white groups.[46] Sommers conducted an experiment that instructed white participants that they would discuss either a race-neutral topic—gay marriage—or a race-relevant topic—affirmative action—in either an all-

40. Samuel R. Sommers, *On Racial Diversity and Group Decision Making: Identifying Multiple Effects of Racial Composition on Jury Deliberations*, 90 J. PERSONALITY & SOC. PSYCHOL. 597, 607 (2006) (finding that diverse groups were less likely to deem the black defendant guilty than all-white groups).

41. *Id.* at 604-06 (finding that diverse groups discussed more facts, evidence that was missing, and race-related concepts than all-white groups).

42. *Id.* at 603.

43. *Id.*

44. Samuel R. Sommers et al., *Cognitive Effects of Racial Diversity: White Individuals' Information Processing in Heterogeneous Groups*, 44 J. EXPERIMENTAL SOC. PSYCHOL. 1129, 1133 (2008). Word-stem completion tasks are those in which a participant is given the beginning of a word, followed by a certain number of blank spaces, and is asked to complete the word. *See, e.g.*, Claude M. Steele & Joshua Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOC. PSYCHOL. 797, 803 (1995). Completing a stem-completion task with race-related words means that an individual fills in an ambiguous stem with a term related to race. *See id.* For example, "_ _ ACK" could be filled out in multiple ways, such as SMACK, which is race-neutral, or BLACK, which is race-related. *Id.* For more examples, see *id.*

45. Sommers, *supra* note 40 at 604-06.

46. Sommers et al., *supra* note 44, at 1133.

white or racially diverse group.[47] The participants never actually discussed the topic in a group, but were led to believe they would.[48] Before reading background material on the subject, but after knowing the topic to which they were assigned, participants individually completed a word-stem completion task.[49] Participants then read background material on their topic and were given a reading comprehension test.[50]

Sommers found that the white participants expecting to discuss a race-relevant topic with a racially diverse group completed word-stems with more racial terms, suggesting that race was more accessible[51] and salient for them. Sommers also found that this increased salience led to better performance on the reading comprehension test.[52] In other words, the influence of racial diversity on reading comprehension is mediated by race salience; race salience explains why those expecting to interact in racially diverse groups process information better.[53] In addition, although those expecting to discuss a race-relevant topic completed word-stems with more race-related terms, they completed fewer word-stems with stereotypical terms – such as "poor," "drugs," and "crime"[54] – suggesting that this race salience suppressed their biases. This finding supports the aversive racism theory framework in that race salience leads to both better information processing for whites (as measured by better performance on the reading comprehension test) and the suppression of biases. In the legal context, discussing the guilt of a black defendant with a racially diverse jury is analogous to discussing a race-relevant topic, which should lead to better information processing and fewer biases among white jurors.

The juries in another of Sommers's studies, though racially diverse, were majority-white, so the results of this study seem to be in tension with the results of the majority-white juries in the study of Dolores Perez and others. Sommers's study found that whites in heterogeneous majority-white juries were nonbiased,[55] whereas Perez's study found that whites in heterogeneous majority-white juries were more likely to convict a Hispanic than a white defendant.[56] These results seem contradictory, but two differences between the

47. *Id.* at 1132.
48. *Id.* at 1131.
49. *Id.* at 1133.
50. *Id.*
51. *Id.*
52. *See id.*
53. *Id.*
54. *Id.*
55. Sommers, *supra* note 40, at 604-05.
56. Perez et al., *supra* note 39, at 1256.

studies suggest that they may in fact be compatible. First, Sommers's study did not manipulate defendant race: the defendant was always black.[57] Therefore, one cannot state conclusively that the juries were less biased toward a black defendant than a white defendant; all that can be concluded is that heterogeneous juries are significantly less likely to convict a black defendant than are homogeneous juries. Second, Sommers compared the juries to each other,[58] whereas Perez and colleagues compared the ratings of each Hispanic or white defendant's guilt within each jury, and then compared the differential findings between the juries.[59] Therefore, it is possible that the heterogeneous jury in Sommers's study, although relatively nonbiased against a black defendant as compared to the all-white jury,[60] may still have treated a white defendant more leniently than a black defendant. That result would fit with the finding that heterogeneous majority-white juries are more biased against minority defendants than against white defendants.[61]

3. Race Salience in this Note

Because, as I have argued, implicit race salience likely suppresses white juror biases, the present research makes race salient in two different ways. This approach tests whether empirical data support my hypothesis about implicit race salience, and, if so, whether my hypothesis can be used to combat juror biases in the form of debiasing jury instructions. It also employs the traditional definition of explicit race salience to attempt to replicate past results and to address the second shortcoming of the literature: the confounding factor of interracial crime. I will first outline the conceptualization of implicit race salience in this experiment, followed by the conceptualization of explicit race salience.

57. Sommers, *supra* note 40, at 602.

58. *Id.* at 603-06.

59. Perez et al., *supra* note 39, at 1256.

60. Sommers, *supra* note 40, at 603-04.

61. Perez et al., *supra* note 39, at 1256. Note that the two studies are not perfectly comparable because white jurors may treat Hispanic individuals differently from black individuals. The aversive racism theory framework is typically applied to whites' interactions with blacks because of whites' guilt over historical racism against blacks; whites compensate by attempting to treat blacks fairly. Thus, the result that heterogeneous majority-white juries are more lenient toward white than Hispanic defendants may not generalize to heterogeneous majority-white juries faced with white defendants in one condition and black defendants in another.

a. Conceptualization of Implicit Race Salience

This Note expands on the idea that race salience can function implicitly and lead to aversive racists' equal judgments of guilt for white and black defendants. Both Skolnick and Shaw's study and the jury studies demonstrate that race need not be explicitly invoked in a trial for it to be salient. In these studies, whites are implicitly aware of race, and this implicit awareness reminds them of their egalitarianism, a process that in turn leads to the suppression of their biases. Expanding the definition of race salience not only strengthens support for aversive racism theory by suggesting that Skolnick and Shaw's study may support the theory but also suggests that explicit race salience is not necessary to decrease juror biases. In both the traditional definition of race salience as explicit and the expanded implicit definition, salience reduces biases because it reminds jurors of their egalitarian views. So egalitarianism, regardless of the definition employed, seems to be the driving force for reducing biases. Explicit and implicit race salience are sufficient, but not necessary, to reduce biases; biases are reduced even when one or both types of salience are absent, whereas egalitarianism is necessary to reduce biases in both instances.

This conceptualization motivates the jury instructions derived from aversive racism theory in this Note: instructions that prime egalitarianism, even without mentioning race, should decrease juror biases. It is important to trigger egalitarianism directly, as not every trial has implicit or explicit race salience. Therefore, jury instructions that are derived from egalitarianism should lower biases against black defendants, even if the jury is homogenously white and there is no implicit societal race salience and no explicit race salience. Such instructions would be a potentially powerful tool for reducing prejudice.

b. Conceptualization of Explicit Race Salience

In addition to basing jury instructions on egalitarianism, the study also employs the standard definition of explicit race salience to (1) replicate prior findings and (2) determine whether egalitarianism and explicit race salience are additive, such that having both egalitarianism and explicit race salience might decrease biases more than either does on its own. The manipulation of explicit race salience draws on the traditional definition of race salience, in which race is a central element of the crime,[62] such that race is a central element of the crime or race had motivated the crime itself.[63] In line with this definition, in this Note, race was mentioned during the commission of the crime by a by-

62. *See* Sommers & Ellsworth, *supra* note 13, at 608.

63. Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1369, 1373.

stander complaining about an increase in crimes committed by blacks. In this sense, race was featured prominently in the description of the crime. However, this approach departs from the traditional definition of explicit race salience. Traditionally, race motivated the crime, such that the *defendant* or *victim* made racist remarks.[64] In this study, race is not the impetus for the crime. Rather, a neutral *bystander* mentions race. This conceptualization of race salience was chosen for two reasons: (1) to eliminate an alternative explanation for any possible results, namely the idea that participants' judgments reflected hostility toward the black victim/defendant as a function of ingroup pride rather than prejudice against the black victim/defendant; and (2) to eliminate the alternative explanation that an interaction between the victim and defendant's race produced the results rather than the defendant's race alone.

One problem with prior studies is that, in the crime scenarios, racial remarks were exchanged between the victim and the defendant, and the victim was either white or black and the defendant was the opposite race.[65] If this Note conceptualized race salience this way and bias were found against the black defendant, it could plausibly be explained by social identity theory. It is quite possible that, for example, white participants might judge black defendants more harshly than white defendants because they feel their self-worth as white individuals is affronted when a black victim's racial insults cause a white defendant to commit a crime, or when a black defendant yells racial insults at a white victim. In other words, white participants may excuse the white defendant's actions based on their anger toward the black victim for inciting violence through racial slurs, or they may want to punish the black defendant more harshly for being racist toward a white victim. In this respect, white jurors may see the black victim's or defendant's racist comments against whites as personally attacking them, thereby affecting their guilt judgments. In fact, this is consistent with social identity theory. Whites may be more lenient toward white defendants because of ingroup pride; the black victim's or defendant's racist comments may wound their self-esteem, leading to comparatively lenient judgments of the white defendant as a means of raising ingroup status and restoring self-esteem.

This is just one example of how both aversive racism theory and social identity theory can plausibly explain why white jurors are more likely to find black defendants guilty than white defendants. Participants may be more lenient toward a white defendant because the black victim threatened their ingroup self-esteem through his racist comments, or they may be harsher toward a black defendant because of implicit prejudice. In other words, although many

64. Sommers & Ellsworth, *supra* note 13, at 603.

65. *E.g.*, Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1372-73.

past studies have shown that explicit race salience leads to the suppression of biases, it may lead to the suppression of biases only when a crime is interracial. Therefore, whether explicit race salience itself leads to the suppression of biases is unknown, and this Note was designed to test this idea. To minimize the impact of the confounding variable of an interracial crime, this Note employed a bystander who mentions race so that participants' judgments of the defendant would not be affected by any animosity toward the victim. Further, if racial hostility motivated the defendant to commit the crime, then the confounding variable of an interracial crime would remain, as it would be unusual for a defendant to commit a racially motivated crime against someone of his own race. By keeping the racial element out of the crime, the racial identity of the victim can be concealed. As a result, any difference in guilt judgments can be attributed to the race of the defendant rather than to any sort of interaction between the defendant's race and the victim's race.

Before turning to social identity theory, one last component of aversive racism theory will be explained: aversive racists will only exhibit bias if the evidence against the defendant is ambiguous.

4. Ambiguity: Why a Lack of Race Salience Causes Juror Biases

The studies discussed in this Note suggest that when race is not explicit, white jurors are not aware of their biases and hence do not try to suppress them, which results in biased decision making. Significantly, however, a lack of race salience seems to result in biases only when evidence concerning the defendant's guilt is ambiguous.[66] In one study, for example, the eyewitness's credibility was suspect, and there was doubt as to the defendant's whereabouts when the murder took place, so his guilt was uncertain.[67] According to aversive racism theory, when there are "nonracial factors in interracial situations, whites may discriminate against blacks and still perceive themselves as being nonprejudiced and egalitarian."[68] In addition to the legal context, this effect has been documented in a variety of contexts.[69] In the face of conflicting evidence (a

66. *E.g.*, Ugwuegbu, *supra* note 2, at 139-40.

67. Skolnick & Shaw, *supra* note 33, at 507-08.

68. Gaertner & Dovidio, *supra* note 8, at 73.

69. Christopher L. Aberson & Tara E. Ettlin, *The Aversive Racism Paradigm and Responses Favoring African Americans: Meta-Analytic Evidence of Two Types of Favoritism*, 17 SOC. JUST. RES. 25, 33, 35 (2004) (conducting a meta-analysis of various studies, including those analyzing helping behavior, medical patient relations, and employment hiring decisions, and finding that when norms were ambiguous, whites favored whites over blacks, but this did not occur when there was no ambiguity). A meta-analysis is an aggregation of multiple studies to discern broad patterns in the results. *See also* John F. Dovidio & Samuel L. Gaertner, *Aversive*

nonracial factor), whites rely more strongly on evidence that incriminates a black defendant, using the ambiguity of the evidence to justify to themselves that they are not being racist. If evidence of innocence were strong, they would not be able to adequately rationalize their bias and would have to suppress this bias. In this respect, whites do not always show bias against blacks; bias is activated only when evidence is ambiguous.

Strength of evidence is a good proxy for ambiguity, as evidence that is questionable is interpreted as ambiguous. Most studies that have found white juror bias have presented ambiguous evidence against the defendant, but these studies did not manipulate the strength of evidence, [70] so one cannot compare across conditions of strong, weak, and ambiguous evidence to determine whether biases only occur when evidence is ambiguous. Fortunately, one study did just that: the evidence was nonexistent, strong, or marginal/ambiguous depending on the condition.[71] The study found that white jurors rated a black defendant's and white defendant's culpability equally if the evidence was strong or weak; when evidence was marginal/ambiguous, on the other hand, white jurors rated the black defendant as more culpable than the white defendant.[72] This study provides empirical support for the idea that biases emerge only when evidence is ambiguous. White jurors likely attribute their harsher treatment of blacks to nonracial factors; for example, they may focus more on evidence that incriminates the black defendant than on evidence that exonerates him. When evidence is strong or weak and hence largely points in one direction, jurors cannot use nonracial factors to justify a biased decision. As a result, for this Note, evidence was purposefully ambiguous in order to test aversive racism theory.

As an added bonus, having ambiguous evidence more accurately mirrors real-life trials. In practice, if evidence against the defendant is strong, the defense attorney will often seek a plea bargain, and if evidence is weak, the prosecutor will often seek a plea bargain. Generally, many of the cases that go to trial are those in which both the prosecutor and defense attorney believe that the evidence is ambiguous.

In sum, the aversive racism theory framework is powerful. It explains why explicit and implicit race salience lead white jurors to treat white and black de-

---

Racism and Selection Decisions: 1989 and 1999, 11 PSYCHOL. SCI. 315, 317 (2000) (finding that when candidate qualifications were ambiguous, black candidates were recommended for a job significantly less strongly than white candidates).

70. *See, e.g.*, Cohn et al., *supra* note 2, at 1958-59; Sommers & Ellsworth, *Race in the Courtroom*, *supra* note 2, at 1368-69, 1372-73; Sommers & Ellsworth, *White Juror Bias*, *supra* note 2, at 214-15.

71. Ugwuegbu, *supra* note 2, at 137.

72. *Id.* at 139.

fendants equally. When race is salient, whether explicitly or implicitly, whites attempt to compensate for their implicit negative feelings toward blacks by suppressing their biases. When race is not salient, and evidence is ambiguous, white jurors' implicit biases emerge, leading to leniency toward white over black defendants. However, social identity theory can just as easily explain white juror biases.

B. Social Identity Theory

As previously mentioned, most scholars have attributed juror biases to aversive racism theory. This Note contends, however, that the results of their studies are also compatible with social identity theory, and therefore it is crucial to test whether social identity theory or aversive racism theory better explains juror biases. Teasing apart these two explanations will not only improve our theoretical understanding of the psychological origins of juror biases, but the two explanations also lead to divergent implications for how to combat biases in the courtroom.

According to social identity theory, individuals express favoritism toward ingroup members, "satisfying their need for positive self-esteem"; they want their ingroup members to be treated well so as to reflect positively on them.[73] In accordance with social identity theory, it is possible that white jurors favor their own race because they wish to increase their own self-esteem. However, one could also argue that, rather than white ingroup favoritism, white jurors' leniency toward white defendants may actually be due to jurors' racially prejudiced stereotyping of blacks; this is what aversive racism theory would predict. However, a phenomenon called the "black sheep effect" calls into question the possibility that aversive racism theory explains juror biases. According to this effect, individuals actually *disfavor* their ingroup under certain conditions. When ingroup members act in an objectionable manner, they are often perceived more negatively than outgroup members who act in similar ways.[74] Ingroup members separate themselves from the individual in order to maintain a positive group image.[75] To test the black sheep effect, Jan-Willem van Prooijen conducted a study that manipulated guilt probability (certain, uncertain) and

73. Miles Hewstone et al., *Intergroup Bias*, 53 ANN. REV. PSYCHOL. 575, 580 (2002).

74. José M. Marques et al., *The "Black Sheep Effect": Extremity of Judgments Toward Ingroup Members as a Function of Group Identification*, 18 EUR. J. SOC. PSYCHOL. 1, 12 (1988).

75. *See id.* at 5.

social categorization (ingroup, outgroup).[76] The results show that guilt probability moderates the ingroup bias effect: participants were more punitive toward ingroup members when the guilt of the ingroup member was certain, but were more punitive toward outgroup members when guilt was uncertain.[77]

The fact that ingroup jurors are more likely to convict an ingroup defendant when evidence is strong underscores the importance of social identity theory in guilt judgments. If racial stereotypes alone drove jurors' decisions, then jurors would not be more punitive toward ingroup members when evidence was strong. Social identity theory explains the black sheep effect well: jurors try to distance a clearly guilty ingroup member as an outlier by acting more punitively toward him. By designating the guilty ingroup member as an outcast, ingroup members seek to maintain a positive image of the rest of the ingroup and by extension preserve their self-esteem.[78]

To summarize, whites use ingroup bias to enhance their group's status as long as the evidence is ambiguous, or at least does not strongly incriminate an ingroup member. Although it seems likely that social identity theory explains this phenomenon, it is only a post hoc explanation. Research suggests a few ways to evaluate whether social identity theory explains decision making by testing the link between ingroup bias and self-esteem. If an individual's need for self-esteem fuels ingroup bias, then bias toward the ingroup should increase self-esteem, or low self-esteem should cause ingroup bias.[79] A meta-analysis has supported the idea that ingroup bias increases self-esteem,[80] but little evidence has supported the idea that low self-esteem leads to ingroup bias.[81]

Considering the fact that ingroup bias increases self-esteem, if self-esteem is sufficiently high before jurors are asked to make a decision regarding guilt, then jurors will not be motivated to rely on ingroup bias. Indeed, social identi-

76. Jan-Willem van Prooijen, *Retributive Reactions to Suspected Offenders: The Importance of Social Categorizations and Guilt Probability*, 32 PERSONALITY & SOC. PSYCHOL. BULL. 715, 718, 720, 723 (2006).

77. *Id.* at 725.

78. Van Prooijen's study did not specifically investigate jurors' racial biases. It investigated bias due to national origin and other affiliations, *id.* at 718, 720, 723, suggesting the need to replicate the effect with whites and blacks.

79. *See* Hewstone et al., *supra* note 73, at 580.

80. P.J. Oakes & J.C. Turner, *Social Categorization and Intergroup Behaviour: Does Minimal Intergroup Discrimination Make Social Identity Theory More Positive?*, 10 EUR. J. SOC. PSYCHOL. 295, 299-300 (1980) (finding that intergroup discrimination increases self-esteem and dismissing alternative hypotheses).

81. Hewstone et al., *supra* note 73, at 580.

ty theory implicates temporary self-esteem.[82] A meta-analysis found support for the hypothesis that temporary self-esteem, but not long-term self-esteem, is related to discrimination against outgroups.[83] The researchers suggest that low temporary self-esteem can lead to discrimination against outgroups in order to increase this self-esteem.[84] This Note tests this inference through jury instructions derived from social identity theory. If whites make decisions that discriminate against an outgroup to increase their temporary self-esteem, then jury instructions that preemptively and temporarily increase jurors' feelings of self-worth should reduce this bias. I term these instructions self-affirming jury instructions.

In the self-affirming instructions, participants were instructed to think about their positive attributes in order to increase their self-esteem.[85] Though studies generally induce self-affirmation through writing exercises rather than thinking exercises,[86] some have asked participants "to use imagery techniques or think about their positive qualities."[87] I decided to pursue a thinking exercise to temporarily raise self-esteem rather than a writing exercise for three reasons. First, prior research has revealed that thinking exercises produce the intended effect by temporarily altering mood.[88] Second, if I had induced self-affirmation through a writing task, the manipulation would not be contained within the jury instructions, and this would risk introducing a confounding variable. Third, a thinking exercise requires less deviation from the current judicial procedures regarding jury instructions. As a result, judges may be more willing to insert self-affirmation elements into jury instructions than to have jurors brainstorm and write about their positive attributes in the middle of jury instructions.

---

82. *See* Mark Rubin & Miles Hewstone, *Social Identity Theory's Self-Esteem Hypothesis: A Review and Some Suggestions for Clarification*, 2 PERSONALITY & SOC. PSYCHOL. REV. 40, 42 (1998) (discussing the distinction between "trait" self-esteem, which is more permanent, and "state" self-esteem, which is temporary).

83. *See id.* at 56-57.

84. *See id.* at 58.

85. *See* discussion *infra* Part IV.B.

86. *See, e.g.*, Amy McQueen & William M.P. Klein, *Experimental Manipulations of Self-Affirmation: A Systematic Review*, 5 SELF & IDENTITY 289, 295-97 (2006) (discussing the types of manipulations used in self-affirmation studies: writing tasks in which participants describe positive experiences, value essays in which participants write about values important to them, and value scales in which participants select values important to them from a list).

87. *Id.* at 295.

88. Joanne V. Wood et al., *This Mood Is Familiar and I Don't Deserve To Feel Better Anyway: Mechanisms Underlying Self-Esteem Differences in Motivation To Repair Sad Moods*, 96 J. PERSONALITY & SOC. PSYCHOL. 363, 366 (2009).

III. HYPOTHESES

A. Aversive Racism Theory Versus Social Identity Theory

Both aversive racism theory and social identity theory can plausibly explain previous findings about biases when evidence is ambiguous; both theories rely on the idea that evidence must be ambiguous for biases to emerge. More specifically, according to social identity theory, evidence must be ambiguous and cannot be strongly incriminating, or else jurors would show a black sheep effect and act more punitively toward the ingroup defendant. According to aversive racism theory, evidence must be ambiguous so that whites have a nonracial excuse – the ambiguity of the evidence – to rationalize their relative harshness toward black defendants. Based on prior research and the theoretical underpinnings of aversive racism theory and social identity theory, I formulated the following set of predictions for each theory. These hypotheses are confined to each theory and do not relate to the issue of which theory is more likely to explain white juror biases.

If aversive racism theory explains juror biases, then participants should be more likely to find black defendants guilty when explicit race salience is low as opposed to high and when there are no jury instructions as compared to when there are egalitarian instructions. Further, one of the deficiencies in the current literature is that participants' levels of implicit and explicit racism were not measured. This is a significant gap in the literature because aversive racism theory relies on whites' possession of high implicit and low explicit racism, but there are also whites that have both high implicit and high explicit racism (true racists) and those who have both low implicit and low explicit racism (non-racists). Therefore, past results cannot accurately be attributed to aversive racism theory because those tested were not necessarily aversive racists. As a result, this Note measured participants' implicit and explicit racism scores so that participants could be split into three groups: true racists, aversive racists, and non-racists.

If aversive racism theory explains white juror biases, then the level of explicit salience and whether there are egalitarian instructions should affect only aversive racists. True racists will be racist even when explicit salience is high and there are egalitarian instructions, because they do not have egalitarian views due to their high explicit racism. Non-racists do not have these biases, so even when explicit salience is low and there are no instructions, they should not exhibit biases. This Note also examines whether the combination of egalitarian instructions and high explicit race salience might decrease judgments of guilt by aversive racists more than either factor on its own.

Social identity theory would predict the following set of results. First, the results should be the opposite of those for aversive racism theory; namely, par-

ticipants should judge the black defendant to be guiltier when explicit salience is high than when it is low, and this should not depend on racism level, as social identity theory applies to all whites.[89] The more salient race is, the more one identifies with one's ingroup and the more likely one is to derogate an outgroup member in order to increase self-esteem. In addition, the black defendant should be judged to be guiltier when there are no instructions compared to the situation in which there are self-affirming instructions. Self-affirming instructions should also reduce any differences in guilt based on explicit race salience by increasing participants' sense of self-worth so that they will not need to discriminate against the black defendant regardless of whether salience is high or low.

B. Procedural Justice

In addition to testing whether theories that specifically deal with race can decrease biases, this Note examines whether procedural justice, a race-neutral theory, can combat biases. If so, this theory could eliminate the need to make race salient in order to decrease racism.

Procedural justice, as noted above, suggests that when individuals experience fair decision-making procedures, they view the law as more legitimate.[90] This legitimacy then leads to compliance with legal rules.[91] Procedural justice is defined by four issues,[92] all of which shaped the jury instructions in this study. First, individuals want to have a voice and share their side of a story.[93] In line

89. Because social identity theory is based on ingroup pride, all members of the ingroup have the motivation to raise the status of the ingroup, and this does not depend on racism level. Thus, unlike aversive racism theory, social identity theory does not predict that levels of implicit and explicit racism will differentially affect guilt judgments. In addition, the race-salience prediction seems to contradict past research showing that high race salience leads to equal judgments of guilt for white and black defendants. *See* sources cited *supra* note 2. However, as mentioned earlier, the results were confounded by the interracial nature of the crimes committed, so the issue of whether high race salience leads to equalization of judgments in accordance with aversive racism theory is unclear. *See supra* Part II.A.3. Therefore, this Note eliminated this confounding variable to determine whether high race salience leads to equal guilt judgments (in line with aversive racism theory), or whether, without an interracial crime, high race salience leads to ingroup identification and biased judgments (in line with social identity theory).

90. *See* TOM R. TYLER, WHY PEOPLE OBEY THE LAW 5 (2006).

91. Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283, 306 (2003).

92. Tom Tyler et al., *What Motivates Adherence to Medical Recommendations? The Procedural Justice Approach to Gaining Deference in the Medical Arena*, 8 REG. & GOVERNANCE 350, 351 (2014).

93. *Id.*

with this idea, the jury instructions used to test this theory told participants to consider all facts. This comports with the idea of voice, as it allowed both sides' stories, the defense's and the prosecution's, to be heard. Second, procedural justice requires unbiased decision makers.[94] For instance, studies have shown that the police are viewed as more legitimate if they exercise authority through fair and impartial means.[95] The instructions primed jurors' sense of impartiality to encourage them to view the juror decision-making process as more legitimate, and as a result, to be unbiased in determining guilt or innocence. Third, individuals want to be treated with civility and dignity.[96] And fourth, individuals want their authority figures to be trustworthy.[97] In line with these final two issues, the jury instructions primed the importance of treating the defendant with respect. It was predicted that jury instructions based on procedural justice would lower biases for all participants, as participants would view their decision-making process as fair and, in turn, comply with the ideal of impartiality in juror decision making.

Research suggests that procedural justice effects can produce long-lasting changes in behavior. For example, one group of researchers found that the perceived fairness of a mediated session predicted compliance with the mediated agreement up to eight months later.[98] In this study, pairs participated in dispute mediation.[99] During the mediation, the mediator told the participants that he would be neutral and that the disputants would get to tell their sides of the story.[100] After the mediation, participants were asked questions that tapped into the construct of procedural justice. Namely, they were asked how fair the mediation procedure was, whether they thought their voice was heard, and whether the mediator understood what they had to say.[101] As procedural justice theory would predict, when respondents viewed the mediation as fair, they were more likely to comply with the terms of the mediated agreement, accord-

94. *Id.*

95. *See* Jonathan Jackson et al., *Why Do People Comply with the Law? Legitimacy and the Influence of Legal Institutions*, 52 BRIT. J. CRIMINOLOGY 1051, 1060 (2012); *see also* Sunshine & Tyler, *supra* note 25, at 535 (stating that procedural justice in policing causes the police to be viewed as legitimate).

96. Tom Tyler et al., *supra* note 92, at 351.

97. *Id.*

98. Dean G. Pruitt et al., *Long-Term Success in Mediation*, 17 LAW & HUM. BEHAV. 313, 324-26 (1993).

99. *Id.* at 316.

100. *Id.* at 317.

101. *Id.* at 318-19.

ing to claimants.[102] Similarly, if jury instructions primed fairness and individuals view the juror decision-making process as fair, participants should comply with the process requiring them to render unbiased verdicts.

C. *Timing of Instructions*

This Note also examines when debiasing instructions have the greatest effect in reducing biases. Differences in guilt judgment based on timing will help suggest when debiasing instructions should be given to juries in actual trials. Currently, juries in criminal trials generally receive the bulk of jury instructions after they hear evidence.[103] At that point, jurors have already formed opinions about the case,[104] and it is very difficult to change their perceptions. Instructions presented before evidence are likely more effective at combatting biases than instructions presented after evidence for two reasons.

First, instructions presented before evidence will help jurors to focus on and recall relevant evidence rather than irrelevant evidence[105] such as race. Second, placing jury instructions before the presentation of evidence can provide jurors with a framework for evaluating evidence. Research has shown that individuals are more able to recognize primed information when they have an organizational schema that provides them with context to evaluate the evidence.[106] Pre-evidence instructions would provide the jury with an organizational framework for evaluating evidence. If the instructions are also debiasing, they will prime jurors to organize the evidence according to legal principles rather than personal biases.

Despite theoretical support for pre-evidence instructions,[107] empirical research has been mixed. On the one hand, some research has found that giving jury instructions before rather than after the evidence results in better recall of

---

102. *Id.* at 324.

103. *E.g.*, Christian Sheehan, *Making the Jurors the "Experts": The Case for Eyewitness Identification Jury Instructions*, 52 B.C. L. REV. 651, 681-82 (2011). *But see, e.g.*, ARIZ. R. CRIM. P. 18.6(c) ("Immediately after the jury is sworn, the court shall instruct the jury concerning its duties, its conduct . . . and the elementary legal principles that will govern the proceeding.").

104. *See* HARRY KALVEN, JR. & HANS ZEISEL, THE AMERICAN JURY 104-17, 375-80 (1966).

105. *E.g.*, Amiram Elwork et al., *Juridic Decisions: In Ignorance of the Law or in Light of It?*, 1 LAW & HUM. BEHAV. 163, 177 (1977).

106. *E.g.*, John D. Bransford & Marcia K. Johnson, *Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall*, 11 J. VERBAL LEARNING & VERBAL BEHAV. 717, 724 (1972).

107. *See supra* text accompanying notes 105-106.

the instructions,[108] lower conviction rates,[109] deferral of determining a defendant's guilt and thus not jumping to a conclusion until after all the evidence has been presented,[110] and appropriate distinctions among plaintiffs who have different levels of deservingness in a civil trial.[111] Yet other research has found that presenting the jury instructions earlier does not affect verdicts,[112] interpretations of the reasonable doubt standard,[113] or evaluations of the strength of evidence.[114]

Although the empirical results are mixed, these studies examined legally substantive jury instructions, or instructions that explain legal standards, which may be less comprehensible than debiasing jury instructions. Many studies have shown that individuals have difficulty comprehending substantive legal instructions,[115] whereas the debiasing instructions in this Note do not contain legal standards[116] and are arguably simpler and easier to understand. Timing should have less of an effect if jurors do not understand the instructions, as they cannot form a framework within which to evaluate evidence regardless of when the instructions are presented. Because the debiasing instructions do not involve complex legal principles, timing should have a stronger effect: the jurors will be able to understand the instructions, and therefore they

108. Larry Heuer & Steven D. Penrod, *Instructing Jurors: A Field Experiment with Written and Preliminary Instructions*, 13 LAW & HUM. BEHAV. 409, 424-25 (1989) (reporting a marginally significant effect for recall of instructions in criminal, but not civil, trials).

109. One study suggests that preliminary instructions affect judgments of guilt, whereas post-evidence instructions have no effect on guilt judgments, functioning as if no instructions had been given. *See* Saul M. Kassin & Lawrence S. Wrightsman, *On the Requirements of Proof: The Timing of Judicial Instruction and Mock Juror Verdicts*, 37 J. PERSONALITY & SOC. PSYCHOL. 1877, 1882 (1979) (finding that preliminary instructions led to forty-one percent guilty verdicts whereas no instructions and post-evidence instructions both led to eighty-two percent guilty verdicts).

110. Vicki L. Smith, *Impact of Pretrial Instruction on Jurors' Information Processing and Decision Making*, 76 J. APPLIED PSYCHOL. 220, 225 (1991).

111. Lynne ForsterLee et al., *Juror Competence in Civil Trials: Effects of Preinstruction and Evidence Technicality*, 78 J. APPLIED PSYCHOL. 14, 19 (1993).

112. Donna Cruse & Beverly A. Browne, *Reasoning in a Jury Trial: The Influence of Instructions*, 114 J. GEN. PSYCHOL. 129, 131 (1987); Smith, *supra* note 110, at 225.

113. Kassin & Wrightsman, *supra* note 109, at 1881.

114. *Id.*

115. *See, e.g.*, V. Gordon Rose & James R.P. Ogloff, *Evaluating the Comprehensibility of Jury Instructions: A Method and an Example*, 25 LAW & HUM. BEHAV. 409, 427 (2001); *see also* Amy E. Smith & Craig Haney, *Getting to the Point: Attempting To Improve Juror Comprehension of Capital Penalty Phase Instructions*, 35 LAW & HUM. BEHAV. 339, 346 (2011) (finding that on average, participants answered more than half the questions regarding the categorization of aggravating and mitigating factors incorrectly).

116. *See* discussion *infra* Part IV.B.

can use the debiasing framework as they evaluate evidence. Finally, the mixed results of timing for legally substantive instructions suggest that timing has at least some effect even when comprehensibility is low, or else some studies would not have found a beneficial effect from pre-evidence instructions. Therefore, more comprehensible instructions should lead to results that provide stronger support for the timing hypothesis. It was thus predicted that debiasing instructions presented before evidence would decrease guilt judgments of the black defendant more than debiasing instructions presented after evidence.

IV. METHOD

A. Participants

Four hundred and twelve Amazon Mechanical Turk[117] participants completed the study, 175 of whom were male.[118] There were 26 Asian/Asian Americans, 30 Latino/Hispanic Americans, 35 Black/African Americans, 16 Native/American Indians, 333 White/European Americans, and 6 Other participants.[119] They ranged in age from 18 to 75, and the median age was 32. Participants were restricted to those whose location was the United States and

---

117. Amazon Mechanical Turk is a website where social science researchers post experiments and participants complete them in exchange for pay. Though relatively new, it is considered "a viable alternative for data collection." Gabriele Paolacci et al., *Running Experiments on Amazon Mechanical Turk*, 5 JUDGMENT & DECISION MAKING 411, 417 (2010). Although there is some concern that Mechanical Turk workers have repeat exposure to standard study measures and thus respond in a more accurate manner, *see, e.g.*, Jesse Chandler et al., *Nonnaïvete Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers*, 46 BEHAV. RES. METHODS 112, 120 (2013), this concern is not likely to be a limitation of this experiment. There are very few Mechanical Turk studies that use the Implicit Association Test (IAT) because of the difficulty of measuring reaction time on Qualtrics, so it is unlikely that respondents have been repeatedly exposed to these measures. Further, participants cannot control their responses on the IAT, *see* discussion *infra* Part IV.C, which would prevent them from responding in a more accurate manner even if they had previously been exposed to the measure.

118. This gender balance, in which 57.52% of the sample is female, is typical of Amazon Mechanical Turk samples generally, in which 64.85% of the worker population is female. Paolacci et al., *supra* note 117, at 412. In fact, the sample in this Note is even more representative of the general population than the typical Amazon Mechanical Turk sample, and Amazon Mechanical Turk samples are already more representative of the U.S. population than are other internet samples and college undergraduate samples. *Id.* at 414.

119. Some participants selected that they were multiple races, which explains why the numbers for the race demographics add up to more than 412. Because Amazon Mechanical Turk does not allow for restriction of a sample based on race, the sample consisted of multiple races rather than only white participants. See *infra* Part VI.D.1 for a discussion of how this limits the results.

who had an approval rate of 95% or more for previous work assignments.[120] The participants were paid $1.10 for fifteen minutes of participation.

B. Design

The experiment employed a 4 x 2 x 2 x 3 between-subjects design. The independent variables were content of instructions (egalitarian, self-affirming, procedural justice, none); timing of instructions (pre-evidence, post-evidence); explicit race salience (high, low); and racism category of the participant (nonracist, aversive racist, true racist). In all conditions, the defendant was black.

For the independent variable of content of instructions, in all conditions the participants were instructed on the reasonable doubt standard in determining guilt. The instructions consisted of a page of text with two to four paragraphs depending on the condition, and were presented in bold typeface. In the egalitarian condition, participants read jury instructions that primed egalitarian views. Specifically, they were instructed:

> Whatever your verdict may be, it must not rest upon baseless speculations. Nor may it be influenced in any way by bias, prejudice, or sympathy. Research has shown that some individuals have unconscious biases that affect their judgments, so please monitor any such biases you may have and do not let them affect your judgments.

In the self-affirming condition, participants read instructions that affirmed their self-worth. Specifically, they were instructed:

> Our system greatly respects jurors' abilities to evaluate evidence and return just verdicts. Studies have shown that jurors thoughtfully and carefully examine evidence, especially if they think of their own positive attributes and experiences before considering the evidence. In light of this research, please take a moment to think about your own positive attributes and experiences. Now that you have reflected on your own positive attributes and experiences, I have full confidence that you will consider the evidence carefully and render an appropriate decision.

In the procedural justice condition, participants read instructions that primed the importance of using fair procedures. Specifically, they were instructed:

---

120. Before accepting work from a participant, the individual running the study can decide whether to accept or reject the participant's work. *See Amazon Mechanical Turk FAQs*, AMAZON, http://aws.amazon.com/mturk/faqs [http://perma.cc/Y9AN-92CK]. Those with an approval rating of ninety-five percent or more have had their work accepted at least ninety-five percent of the time.

> Whatever your verdict may be, it is important that you use fair procedures in forming your judgments about the case and in trying to reach your verdict. There are several key elements in procedural justice. First, each juror should consider all the evidence presented about the case in an open minded way and not be influenced by bias, prejudice, or sympathy. You should focus on facts, and avoid being swayed by either prejudice or sympathy. Second, when forming your opinions you should fairly consider all of the evidence. Your job is to identify the facts you need to apply the law fairly and reach a fair verdict. Third, when making decisions as a juror it is important that you do your best to do what is fair for the defendant in terms of the facts of the case and the laws in your community. These rules are designed to help guarantee a fair trial, and our law accordingly sets forth serious consequences if the rules are not followed.

In the no instructions condition, participants did not read any jury instructions except for the brief explanation of the reasonable doubt standard that participants in all conditions read.

For the independent variable of timing of instructions, instructions were given either before or after the presentation of evidence.

For the independent variable of explicit race salience, the salience of race in the description of the crime was either high or low. In the high race salience condition, a bystander was heard complaining about an increase in crimes committed by blacks in the neighborhood. In addition, participants saw a picture of the black defendant. In the low race salience condition, race was not mentioned: the bystander was heard complaining about an increase in crime (not specific to race). In addition, there was no picture of the defendant, though participants were told the race of the defendant in the written description of the defendant so that they knew the defendant's race among other information.

For the independent variable of racism category, there were three categories: non-racist, aversive racist, and true racist. Participants were categorized by combining each participant's implicit and explicit racism scores. The implicit racism score had two levels (high and low), which was determined by a median split of the implicit racism data, obtained using an implicit racism measure designed for online research by Jordan LaBouff.[121] The explicit racism score also had two levels (high and low), which was determined by a median split of the explicit racism data obtained using an explicit racism measure, the Modern

121. Jordan LaBouff, A Brief Online Survey-Based Implicit Association Test for Intergroup Attitudes (Jan. 18, 2013) (poster presented at Society for Personality and Social Psychology) (on file with author) (introducing the Qualtrics-based implicit racism measure).

Racism Scale.[122] Then four groups were created based on participants' implicit and explicit scores. Those with both low implicit and low explicit racism scores were labeled non-racists, as they exhibited a lack of racism in both measures. There were 104 participants in this group. Those with high implicit and low explicit racism scores were labeled aversive racists in line with the definition of aversive racism. There were 81 participants in this group. Lastly, those with both high implicit and high explicit racism scores were labeled true racists, as they had high levels of racism in both measures. There were 101 participants in this group. Those who were low in implicit racism and high in explicit racism were excluded from the subsequent analyses, as this combination is not readily explainable by any race theories; it is unusual not to have implicit bias but to be outwardly and explicitly biased. There were 75 participants in this group.[123]

Six dependent variables were measured for each crime: guilt of the defendant, confidence in guilt/innocence judgment, perceived prior record, sentence judgment, recall of evidence, and recognition of evidence. The guilt of the defendant was measured on a seven-point scale (1 = *not at all likely to be guilty* to 7 = *extremely likely to be guilty*), as was confidence in guilt/innocence judgment (1 = *not at all confident* to 7 = *extremely confident*), perceived prior record (1 = *definitely not* to 7 = *definitely*), and sentence judgment (1 = *0 years* to 7 = *5+ years*).[124]

Participants' recall of evidence pertaining to the crime was measured by asking participants to list as much information regarding the crime as they could remember.[125] Information recalled was coded by counting the number of incriminating and exonerating pieces of evidence each participant recorded. The number of times participants mentioned each type of evidence corresponded to their score for that variable; for example, if a participant mentioned three incriminating pieces of evidence and two exonerating pieces of evidence for the crime, the score would be three and two for each type of evidence, respectively. Recognition of the evidence pertaining to the crime consisted of twelve items, six of which had actually been included in the evidence and six of

122. John B. McConahay, *Modern Racism, Ambivalence, and the Modern Racism Scale*, *in* PREJUDICE, DISCRIMINATION, AND RACISM 91, 104 (John F. Dovidio & Samuel L. Gaertner eds., 1986).

123. The groups had roughly equal numbers because implicit and explicit racism scores were determined by a median split of the data.

124. The two measures consisting of guilt of the defendant and participants' confidence in their guilt/innocence judgments were adapted from Galen Bodenhausen. *See* Galen V. Bodenhausen, *Stereotypic Biases in Social Decision Making and Memory: Testing Process Models of Stereotype Use*, 55 J. PERSONALITY & SOC. PSYCHOL. 726, 729 (1988). The confidence judgment was not included in any analyses because it was not related to any of the hypotheses but instead was used to maintain consistency with Bodenhausen's measure.

125. This dependent variable was also adapted from Bodenhausen. *Id.* at 730, 734.

which were foils. For each type of evidence (actual evidence and foil evidence), there were two incriminating statements, two exonerating statements, and two neutral statements. Participants indicated on a seven-point scale whether they thought each of the statements had been included in the crime scenario, where 1 = *definitely not* and 7 = *definitely*. The neutral pieces of evidence were not included in analyses, as they were not related to any hypotheses of the study.

C. *Materials*

The implicit racism measure was a Qualtrics-based Implicit Association Test (IAT) that measured participants' levels of implicit racism and operated as follows. Participants completed three blocks of trials: the first was not of interest, as it was meant to acclimate participants to the task, but the subsequent two were of interest. In the first block, participants viewed two columns running down the page pairing categories together—the left column was flower/pleasant and the right column was insect/unpleasant. To the left of the columns were words that fit into one of the two columns: words that describe flowers (for example, Geranium), words that describe insects (for example, Centipede), words that describe pleasant things (for example, love), and words that describe unpleasant things (for example, vomit). Participants had to click the radio button in the correct column to categorize the terms on the left into their appropriate groups. Participants were instructed to begin at the top and run down the left side of the page when categorizing the terms. Participants were given thirty-five seconds to complete the block, and then the page automatically advanced when time was up. This time was chosen to make it impossible to categorize all terms within the timeframe, causing some terms to be left unpaired.

After this initial introductory block, the blocks of interest followed. Participants were not instructed that only these blocks were of interest. These blocks followed the same procedure as the initial block, but the category pairings changed. In one of the two subsequent blocks the left column was black/pleasant and the right column was white/unpleasant; in the other of the two subsequent blocks the left column was black/unpleasant and the right column was white/pleasant. The presentation of these blocks was randomized to prevent any order effects. The words that described pleasant and unpleasant items were the same as in the first block. Instead of using words to describe black and white, pictures of black and white males and females were used in

order to mirror how a traditional race IAT operates. The faces were selected and matched for age from the Center for Vital Longevity database.[126]

There were 120 trials total, of which 80 were of interest. The trials of interest were those from the second and third blocks. Unfortunately, Qualtrics does not have the ability to track individual item response times, which is how a typical IAT is scored. To circumvent this issue, LaBouff's version uses the total number of correct responses in an extremely limited time to estimate reaction time in a given condition. The trials were scored using the formula [(*X*/*Y*)SQRT(*X*-*Y*)] and the following steps.[127] First, the IAT items were recoded so that correct pairings were coded as 1 and incorrect or no response pairings were coded as 0.[128] Then the number of trials that the participant got correct was computed in both the congruent and incongruent versions.[129] The congruent version occurs when white is paired with pleasant and black is paired with unpleasant, because those pairings are more natural pairings for those who are implicitly racist. The incongruent version occurs when white is paired with unpleasant and black is paired with pleasant. For the formula, *X* is the number correct in whichever condition the participant got more correct (for example, if incongruent correct is greater than congruent correct, then *X* = incongruent correct).[130] *Y* is the number correct in whichever condition the participant got fewer correct (for example, if incongruent correct is greater than congruent correct, then *Y* = congruent correct).[131] These numbers were plugged into the formula [(*X*/*Y*)SQRT(*X*-*Y*)] for each participant. Then the direction of the effect was reversed for participants who had incongruent scores that were greater than their congruent scores.[132] In short, the trials of interest operated like the IAT, so that high implicit racism corresponds with participants who answered more items correctly in the congruent condition than in the incongruent condition.

Explicit racism was measured using the five-point Modern Racism Scale. Participants were asked the degree to which they agreed with each of seven

126. *See* Meredith Minear & Denise C. Park, *A Lifespan Database of Adult Facial Stimuli*, 36 BEHAV. RES. METHODS, INSTRUMENTS & COMPUTERS 630 (2004); Park Aging Mind Lab., *Stimuli*, U. TEX. DALL., http://agingmind.utdallas.edu/facedb [http://perma.cc/5NJJ-YN63].

127. E-mail from Jordan LaBouff, Professor, Univ. of Me., to author (Apr. 22, 2014, 20:41 EST) (on file with author).

128. *Id.*

129. *Id.*

130. *Id.*

131. *Id.*

132. *Id.*

statements where 1 = *Strongly Disagree* and 5 = *Strongly Agree*.[133] This measure was adapted from John McConahay's work.[134] The scale was reliable ($\alpha = .73$).

D. Procedure

Participants were told that they would read and complete a survey about a crime scenario to help with research that was examining how jurors evaluate evidence and determine a defendant's guilt. Participants then read the consent form and gave their consent. They were then told that they would read a crime scenario and answer questions about the scenario. Participants in the pre-evidence instructions condition then read jury instructions. All participants then were presented with a description of the crime committed and information about the suspect.

After reading the description of the crime, participants read thirteen items of evidence pertaining to the crime. Five items were incriminating, five were exonerating, and three were neutral. The evidence items were adapted from Bodenhausen's studies[135] and were randomized to prevent any order effects. After reading the evidence, participants in the post-evidence instructions condition read jury instructions. All participants then answered questions regarding the defendant's guilt and were asked to recall the evidence and perform a recognition task with the evidence. Next the participants completed the Qualtrics-based IAT followed by the Modern Racism Scale. Finally, the participants provided demographic information.

V. RESULTS

A. Descriptive Statistics

Descriptive statistics are presented in Table 1. Any participant who spent less than twenty seconds reading the jury instructions (as measured by a timer embedded in Qualtrics, unknown to participants) was excluded from the analyses. It took, on average, one minute to read the instructions, so those who spent less than twenty seconds reading the instructions likely were not adequately exposed to the independent variable. In total, 50 participants were excluded from analysis, leaving 362 participants whose responses were analyzed.

---

133. One of the items was reverse coded.

134. McConahay, *supra* note 122, at 108.

135. Bodenhausen, *supra* note 124, at 729.

B. Main Results

To test whether participants' guilt judgments, perceived prior record, and sentence judgments differed depending on the content of the jury instructions, explicit race salience, the participant's racism category, and the timing of instructions, I conducted a 4 (content of jury instructions: egalitarian, self-affirming, procedural justice, none) x 2 (explicit race salience: high, low) x 2 (timing of instructions: pre-evidence, post-evidence) x 3 (participant's racism category: non-racist, aversive racist, true racist) analysis of variance (ANOVA).[136] This ANOVA was run for each of the three dependent variables related to guilt: guilt judgment, perceived prior record, and sentence judgment, as well as for the recall and recognition data. There were no significant effects[137] for recognition or recall of data, so these variables were dropped from subsequent analyses.

The general hypothesis in support of aversive racism theory is that only aversive racists would change their guilt judgments depending on the level of explicit race salience and the presence of egalitarian instructions. It was expected that they would find the black defendant to be most guilty when there was low explicit race salience and no egalitarian instructions (either no instructions, self-affirming instructions, or procedural justice instructions). Further, they would find the black defendant less guilty when race was explicitly salient

---

136. An ANOVA measures the total variability in an experiment as a result of the variability between and within groups in order to determine whether there are meaningful differences between groups. *See, e.g.*, Sylvan Wallenstein et al., *Some Statistical Methods Useful in Circulation Research*, 47 CIRCULATION RES. 1, 3 (1980). Although continuous variables are typically analyzed using a multivariate regression, an ANOVA is more appropriate here. Aversive racism theory is predicated on the different categories of racism: racism is seen as a dichotomous, rather than continuous, variable, so to stay true to the theory, I converted the continuous implicit and explicit racism variables into dichotomous racism categories and ran an ANOVA.

137. In research, a significant effect is typically when $p < .05$. *See, e.g.*, Timothy R. Levine & Craig R. Hullett, *Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research*, 28 HUM. COMM. RES. 612, 614 (2002). This technically means that, assuming the null hypothesis is true, or that the variables are unrelated, there is a five percent chance of obtaining the relationship between the variables in the given sample. *Id.* When $p < .05$, the null hypothesis is rejected and results are assumed to be significant. *Id.* When $p < .10$ but $> .05$, the results are assumed to be marginally significant. *P*-values are biased by sample size, however, so another common measure reported is effect size, which is an estimate of how closely two variables are related. *Id.* One such measure of effect size is partial eta squared (η_p^2), *id.* at 616, which, in addition to *p*-values, I report throughout the Note. The higher the value, the larger the effect, and the typical rule of thumb is that .20, .50, and .80 are small, medium, and large effects, respectively. *See, e.g.*, Lee Sechrest & William H. Yeaton, *Magnitudes of Experimental Effects in Social Science Research*, 6 EVALUATION REV. 579, 585 (1982).

or when they read egalitarian instructions. In addition, I wanted to see what would happen when there were both high explicit race salience and egalitarian instructions. This part of the study was exploratory and meant to determine whether explicit race salience and egalitarian instructions are additive and whether both would suppress aversive racists' biases more than either factor alone. It was predicted that true racists and non-racists would have the same ratings of guilt regardless of levels of explicit salience or content of jury instructions. Moreover, overall there should be a main effect of racism category, such that true racists should give the highest ratings of guilt to the black defendant, followed by aversive racists and followed by non-racists.

The general hypothesis in support of social identity theory is that differences in racism category should not affect judgments of guilt. Rather, for all participants, judgments of the guilt of the black defendant should be higher when race is explicitly salient than when it is not explicitly salient. Additionally, participants should rate the defendant as guiltier when there are no self-affirming instructions (either no instructions, egalitarian instructions, or procedural justice instructions) than when there are such instructions. Self-affirming instructions should eliminate any differences in guilt judgment based on explicit race salience.

The general hypothesis in support of procedural justice is that differences in racism category should not affect judgments of guilt. Rather, all participants should rate the defendant as guiltier when there are no procedural justice instructions (either no instructions, egalitarian instructions, or self-affirming instructions) than when there are such instructions.

The general hypothesis in support of timing is that the debiasing instructions, if they work, should be more effective if presented earlier, such that judgments of guilt will be lower when the instructions are presented pre-evidence than when they are presented post-evidence.

1. *Aversive Racism Theory*

The results support aversive racism theory. Because I did not find support for social identity theory and procedural justice, I conducted a power analysis[138]

138. A power analysis determines the probability of correctly rejecting the null hypothesis when it is false. Jacob Cohen, *Statistical Power Analysis*, 1 CURRENT DIRECTIONS PSYCHOL. SCI. 98, 98 (1992). Power is affected by sample size, such that a larger sample is more likely to lead to correctly rejecting a false null hypothesis than a small sample size. *Id.* Therefore, too small of a sample size may result in Type II error, or a false negative. The convention in social psychology is that .80 sets the value for power. *Id.* at 100. This effectively means that in running a post hoc power analysis, one wants the result to be at least .80. If the result is less than .80, the study is underpowered.

to examine how meaningful these null results were. I found that my study was underpowered,[139] which means that I cannot state that the data provide evidence against social identity theory or procedural justice – rather, these theories are just not positively supported by the data.

a. Guilt Judgment

The mean guilt rating that non-racists gave the black defendant was less than the mean guilt rating from aversive racists, which was less than the mean guilt rating from true racists.[140] Post hoc comparisons using Tukey's test[141] revealed that there was a significant difference in guilt judgment between the non-racists and the true racists, $p = .002$, but not between the other groups.[142]

In addition, I performed an exploratory analysis by restricting the sample further. Instead of including participants who spent at least twenty seconds reading the jury instructions, I included only participants who spent at least forty seconds. This exploratory analysis supports aversive racism theory and surprisingly suggests that there is a marginally significant result in which the procedural justice instruction led to increased biases for aversive racists.

In this exploratory analysis, there was a marginally significant interaction between content of instructions and racism category.[143] For the simple effects[144]

---

139. None of the main effects or interactions for guilt judgment approached observed power of .80, except for the main effect of racism category, which had observed power of .96; the three-way interaction between content of instructions, timing, and racism category, which had observed power of .71; and the four-way interaction between all four independent variables, which had observed power of .79. For all other main effects and interactions for guilt judgment, observed power ranged from .06 to .41. None of the main effects or interactions for perceived prior record approached observed power of .80; instead, observed power ranged from .05 to .42. None of the main effects or interactions for sentence judgment approached observed power of .80; instead, observed power ranged from .05 to .61.

140. *See infra* Figure 1 and Table 2. ANOVA result: $F(3, 360) = 4.61$, $p = .004$, $\eta_p^2 = .037$. Means and standard deviations for each racism category: non-racists ($M = 3.13$, $SD = 1.36$), aversive racists ($M = 3.57$, $SD = 1.46$), true racists ($M = 3.87$, $SD = 1.57$).

141. Tukey's test is a post hoc pairwise comparison used when there is a significant main effect but there are more than two levels of a variable. *See, e.g.*, Wallenstein et al., *supra* note 136, at 4 (mentioning that there are five groups and that Tukey's test is appropriate for all pairwise comparisons). When there are more than two levels of a variable, a significant main effect does not reveal which levels differ from each other. Tukey's test does just that: it compares the means within the main effect to determine which means differ from each other. In the sentence accompanying this footnote, Tukey's test revealed that there was only a difference between the non-racists and the true racists, which clarifies the main effect.

142. *See infra* Figure 1 and Table 2.

143. *See infra* Figure 2 and Table 3. ANOVA result: $F(3, 281) = 1.75$, $p = .078$, $\eta_p^2 = .056$.

split by instructions, there was, in the no instructions condition, a significant difference in judgments of guilt among the racism categories.[145] Post hoc comparisons using Tukey's test revealed a significant difference in guilt judgment between the non-racists and the true racists, $p = .013$, such that non-racists who did not read instructions rated the defendant as less guilty than true racists who did not read instructions.[146] There were no differences between the other groups.[147] For the simple effects split by instructions, there was, in the procedural justice instructions condition, a marginally significant difference in judgments of guilt among the racism categories.[148] Post hoc comparisons using Tukey's test revealed a significant difference in guilt judgments between the non-racists and the aversive racists, $p = .04$, such that non-racists who read the procedural justice instructions rated the defendant as less guilty than aversive racists who read the procedural justice instructions.[149] There were no differences between the other groups.[150]

For the simple effects split by racism category, there was, in the non-racist category, a marginally significant difference in judgments of guilt among the instructions conditions.[151] Post hoc comparisons using Tukey's test revealed a marginally significant difference in guilt judgments between the egalitarian instructions and no instructions conditions, $p = .057$, such that non-racists who read the egalitarian instructions rated the defendant as guiltier than non-racists

144. When there is an interaction between at least two variables with at least two levels each, a simple effects test is run to determine where the difference in the interaction lies. Simple effects measure the differences between means within each level of one of the independent variables. Oliver Schabenberger et al., *Collections of Simple Effects and Their Relationship to Main Effects and Interactions in Factorials*, 54 AM. STATISTICIAN 210, 211 (2000). For example, in the simple effects analysis of the marginally significant interaction between the content of instructions and racism category, the data were split by instructions, such that the means of each racism category were compared with each other within each instructions condition. Thus, the means of the guilt judgment of the non-racists, aversive racists, and true racists were compared to each other within the no instructions condition, again within the egalitarian instructions condition, and so forth. The data were then split by racism category, and the means of each instructions condition were then compared within each racism category. Therefore, the means of the guilt judgment in the no instructions, egalitarian instructions, self-affirming instructions, and procedural justice instructions conditions were compared to each other within the non-racists category, the aversive racists category, and so forth. The means that differ significantly in any of these comparisons are then reported.

145. *See infra* Figure 2 and Table 3. ANOVA result: $F(3, 56) = 3.35$, $p = .026$, $\eta_p^2 = .159$.

146. *See infra* Figure 2 and Table 3.

147. *See infra* Figure 2 and Table 3.

148. *See infra* Figure 2 and Table 3. ANOVA result: $F(3, 71) = 2.56$, $p = .062$, $\eta_p^2 = .101$.

149. *See infra* Figure 2 and Table 3.

150. *See infra* Figure 2 and Table 3.

151. *See infra* Figure 2 and Table 3. ANOVA result: $F(3, 81) = 2.73$, $p = .05$, $\eta_p^2 = .095$.

who did not read instructions.[152] There were no differences between the other groups.[153] For the simple effects split by racism category, there was, in the aversive racist category, a significant difference in judgments of guilt among the instructions conditions.[154] Post hoc comparisons using Tukey's test revealed a significant difference in guilt judgments between the egalitarian instructions and procedural justice instructions conditions, $p = .013$, such that aversive racists who read the egalitarian instructions rated the defendant as less guilty than those who read the procedural justice instructions.[155] There were no differences between the other groups.[156]

b. Perceived Prior Record

Non-racists rated the black defendant as least likely to have a prior criminal record, followed by aversive racists; true racists rated the black defendant as most likely to have a prior record.[157] Post hoc comparisons using Tukey's test revealed a marginally significant difference in perceived prior record judgment between the non-racists and the true racists, $p = .088$, but not between the other groups.[158]

c. Sentence Judgment

The mean sentence non-racists gave the black defendant was shorter than the mean sentence from aversive racists, which was shorter than the mean sentence from true racists.[159] Post hoc comparisons using Tukey's test revealed a marginally significant difference in sentence judgment between the non-racists and the true racists, $p = .079$, but not between the other groups.[160]

---

152. *See infra* Figure 2 and Table 3.

153. *See infra* Figure 2 and Table 3.

154. *See infra* Figure 2 and Table 3. ANOVA result: $F(3, 61) = 3.46$, $p = .022$, $\eta_p^2 = .152$.

155. *See infra* Figure 2 and Table 3.

156. *See infra* Figure 2 and Table 3.

157. *See infra* Figure 3 and Table 4. ANOVA result: $F(3, 360) = 2.41$, $p = .067$, $\eta_p^2 = .020$. Note that the result is marginally significant. Means and standard deviations for each racism category: non-racists ($M = 3.30$, $SD = 1.23$), aversive racists ($M = 3.51$, $SD = 1.31$), true racists ($M = 3.72$, $SD = 1.32$).

158. *See infra* Figure 3 and Table 4.

159. *See infra* Figure 4 and Table 5. ANOVA result: $F(3, 360) = 2.28$, $p = .079$, $\eta_p^2 = .019$. Note that the result is marginally significant. Means and standard deviations for each racism category: non-racists ($M = 2.79$, $SD = 1.23$), aversive racists ($M = 2.95$, $SD = 1.37$), true racists ($M = 3.15$, $SD = 1.47$).

160. *See infra* Figure 4 and Table 5.

In addition, there was a three-way interaction among content of instructions, timing of instructions, and explicit race salience.[161] For simple effects, the only significant effect was when the data were split by timing and split again by content of instructions, such that participants' sentence judgments were longer in the low explicit race salience condition than in the high explicit race salience condition, and this difference only occurred when the instructions were egalitarian and presented pre-evidence.[162]

2. Timing of Instructions

The only dependent variable that supported the timing hypothesis was the guilt variable. Participants gave the black defendant a lower guilt rating when instructions were presented pre-evidence than when they were presented post-evidence.[163]

VI. DISCUSSION

This study tested whether jury instructions based on aversive racism theory, social identity theory, or procedural justice would mitigate juror biases against black defendants. In addition to investigating how the content of jury instructions affects judgments of guilt, the study also tested how the timing of jury instructions affects these judgments. Analyses of the data yield support for the timing hypothesis, preliminary support for aversive racism theory, and no support for social identity theory or procedural justice. These results suggest that judges should include debiasing elements derived from aversive racism theory in their jury instructions and that they should present these instructions before the evidence phase of a trial.

A. Support for Timing Hypothesis

In support of the timing hypothesis, participants found the defendant to be less guilty when the debiasing instructions were presented before the evidence as compared to when they were presented after the evidence (though the result was marginally significant).[164] This is a powerful result, as it suggests that

161. *See infra* Figure 5 and Table 6. ANOVA result: $F(2, 360) = 5.27$, $p = .006$, $\eta_p^2 = .029$.

162. *See infra* Figure 5 and Table 6. ANOVA result: $F(1, 52) = 4.30$, $p = .043$, $\eta_p^2 = .078$.

163. *See infra* Figure 6 and Table 7. ANOVA result: $F(1, 360) = 3.10$, $p = .079$, $\eta_p^2 = .009$. Note that the result is marginally significant. Means and standard deviations for each timing condition: pre-evidence ($M = 3.39$, $SD = 1.40$), post-evidence ($M = 3.73$, $SD = 1.54$).

164. *See supra* note 163 and accompanying text.

courts should reconsider current procedures for jury instructions. Currently, judges typically read instructions to the jury after the jurors have heard all the evidence in a trial.[165] According to the results of this study, presenting instructions before evidence leads to lower biases. Therefore, it would be beneficial for judges to present debiasing instructions before jurors hear evidence.

B. Support for Aversive Racism Theory

In support of aversive racism theory, there was a significant main effect of racism category on guilt judgment,[166] and marginally significant main effects on perceived prior record[167] and sentence judgment,[168] such that non-racists found the black defendant to be less guilty than aversive racists, who found the black defendant to be less guilty than true racists.[169] Though the only significant difference was between the non-racists and true racists,[170] all of the ratings trended in the predicted direction.[171] This finding is particularly important, as it adds to the literature by demonstrating that both implicit and explicit racism levels predict judgments of guilt, and each contributes separately to these predictions as shown by the increase in guilt judgments as both implicit and explicit racism increases.

In further support of aversive racism theory, there was a three-way interaction among content of instructions, timing of instructions, and explicit race salience.[172] Participants gave the black defendants a longer sentence when explicit race salience was low than when it was high, but only when egalitarian instructions were presented pre-evidence.[173] This result supports aversive racism theory: it suggests that the combination of explicit race salience and egalitarian instructions can be powerful in reducing biases. It also supports the exploratory hypothesis that explicit race salience and egalitarianism may be additive, such

165. Sheehan, *supra* note 103, at 681-82. I recognize that judges usually also give instructions before evidence is presented, but generally these instructions are not as extensive as the instructions given prior to jury deliberation. In addition, the jury instructions in this Note included not only the burden of proof, but also debiasing elements. The timing hypothesis focused on determining whether, if debiasing instructions are used, they are better used before or after the presentation of evidence.

166. *See supra* note 140 and accompanying text.

167. *See supra* note 157 and accompanying text.

168. *See supra* note 159 and accompanying text.

169. *See supra* notes 140, 157, 159 and accompanying text.

170. *See supra* notes 142, 158, 160 and accompanying text.

171. *See supra* notes 140, 157, 159 and accompanying text.

172. *See supra* note 161 and accompanying text.

173. *See supra* note 162 and accompanying text.

that their combination is more powerful at reducing biases than either alone. In addition, this result provides support for the timing hypothesis, as the debiasing instructions only had their effect when they were presented pre-evidence.

However, aversive racism theory was not fully supported by the data, as it should not be necessary that both egalitarian instructions and explicit race salience be present in order to decrease biases. Both factors should prime egalitarian views, so each on its own should have led to decreased biases. Further, explicit race salience only mattered when there were egalitarian instructions presented pre-evidence,[174] which deviates from past research in which explicit race salience affected judgments whenever race was explicitly salient. However, as previously mentioned, past studies manipulated explicit race salience by including a racial factor as part of the impetus for the crime,[175] thereby creating an interracial crime. Consequently, although explicit race salience mattered in judgments of guilt, it was confounded by a potential interaction between the defendant's and victim's race. Because this Note excluded the victim's race in order to avoid this confounding factor, the data suggest that perhaps the results from past studies have been driven in part by an interaction between the defendant's race and the victim's race rather than by explicit race salience alone.

Additionally, although there was a main effect of racism category on guilt judgment, perceived prior record, and sentence judgment, racism category did not interact with the content of the instructions or explicit race salience for guilt judgments or perceived prior record, as predicted by the theory. Instead, guilt ratings and perceived prior record ratings were the same for aversive racists (and non-racists and true racists) regardless of the content of the instructions and explicit race salience. For the findings to fully support aversive racism theory, the effects of explicit race salience and content of instructions should have been moderated by racism category.

Though racism category did not interact with content of instructions or explicit race salience, an exploratory analysis suggests that this interaction may have occurred for participants who spent at least forty seconds reading the jury instructions. For the primary data analyses I could eliminate only participants who spent less than twenty seconds on the page in order to maintain sufficient statistical power. Eliminating more participants would have excluded too large a proportion of the total participants. As a result, I ran an exploratory analysis in which I included only participants who spent at least forty seconds reading the instructions. This exploratory analysis yielded a significant interaction between racism category and content of instructions.[176] Within this interaction,

174. *See supra* note 162 and accompanying text.

175. *See supra* Part II.A.3.

176. *See supra* note 143 and accompanying text.

there was a marginally significant simple effect[177] that supports aversive racism theory: non-racists rated the black defendant as guiltier in the egalitarian condition than in the no instructions condition.[178] At first glance, this does not support aversive racism theory: non-racists should have low ratings of guilt regardless of whether they are given debiasing instructions. They are non-biased, so any debiasing instructions will not affect their guilt judgments. However, those that were labeled non-racists were not actually non-biased. In fact, they had implicit preferences for blacks, according to the scoring from LaBouff's Qualtrics-based IAT. This finding is unusual, as most IATs show that very few individuals have implicit preferences for blacks; most have no preferences between blacks and whites or else they have a preference for whites over blacks.[179]

Regardless of what caused the implicit preference for blacks, it is significant because the non-racists' results comport with aversive racism theory. Their baseline bias is in the opposite direction as that of aversive racists. If "non-racist" participants are implicitly biased against whites but have low explicit racism, when they read egalitarian instructions they will be reminded of their egalitarian views and will *increase* their guilt ratings for black defendants. Therefore, the finding that non-racists had higher guilt ratings for the black defendant in the egalitarian instructions condition than in the no instructions condition actually supports aversive racism theory.

In addition, as in the primary analyses, there was a trend in the predicted direction, though not significant, in the effect of racism category on guilt ratings, such that in the no instructions condition, non-racists gave the lowest guilt ratings, followed by aversive racists, and followed by true racists.[180] Further, aversive racists had lower guilt ratings in the egalitarian condition than in the procedural justice condition,[181] and this result supports aversive racism theory: the egalitarian instructions reminded them of their desire to be nonbiased, so they decreased their biases in line with this ideal. Also in line with aversive racism theory, true racists had very similar ratings in all instructions conditions[182]: since they are racist and do not have egalitarian views, debiasing instructions should not decrease their biases.

---

177. *See supra* note 151 and accompanying text.

178. *See supra* note 152 and accompanying text.

179. The sample was a mixed race sample, so perhaps any minorities in the sample – particularly blacks – might explain the larger than normal portion of participants that showed an implicit preference for blacks. However, running the analyses without the minority participants produced similar results, so it is unlikely minority participants drove this reverse bias.

180. *See infra* Figure 2 and Table 3.

181. *See supra* note 155 and accompanying text.

182. *See infra* Figure 2 and Table 3.

C. *Lack of Support for Social Identity Theory and Procedural Justice*

The data do not support social identity theory; the self-affirming instructions condition did not decrease biases any more than the other instructions or no instructions conditions.[183] In addition, the explicit race salience data partially supports aversive racism theory, and hence cannot support social identity theory, since aversive racism theory and social identity theory predict opposite results. According to social identity theory, increased salience should lead to increased biases, whereas according to aversive racism theory, increased salience should lead to decreased biases. Thus, because the explicit race salience data supports aversive racism theory, such that higher salience decreased biases, it does not support social identity theory, as higher salience should have increased biases. The data also do not support the procedural justice hypothesis; the procedural justice instructions condition did not decrease biases any more than the other instructions or no instructions conditions. Also, in the exploratory analysis with participants who spent at least forty seconds on the instructions page, procedural justice instructions actually caused aversive racists to increase their guilt ratings.[184]

D. *Limitations and Directions for Future Research*

Although this Note explored many new ideas in the juror decision-making literature, there were limitations to the study. These limitations include the mixed race sample, the lack of comparison to a white defendant, the implicit racism measure employed, and the artificiality of the online setting. Each limitation, and a means to overcome it in future studies, will be addressed in turn.

1. *Restriction of Sample to White Participants*

In the future it would be prudent to restrict the sample to white participants, as aversive racism theory is focused on white juror bias, and social iden-

183. One important caveat is that the self-affirming manipulation may have inadvertently prevented participants from feeling self-affirmed. The self-affirming manipulation *informs* the subjects that "[s]tudies have shown that . . . thinking of their own positive attributes and experiences" was an experience that could be expected to improve the fairness of their decision making. Some studies have shown that self-affirmation eliminates bias better when participants are not aware they are being self-affirmed. *See, e.g.*, David K. Sherman et al., *Affirmed Yet Unaware: Exploring the Role of Awareness in the Process of Self-Affirmation*, 97 J. PERSONALITY & SOC. PSYCHOL. 745, 757 (2009). This effect could potentially explain why the data did not support social identity theory.

184. *See supra* note 149 and accompanying text.

tity theory is based on ingroup preferences, so having multiple races interferes with the results. This may explain why the data did not support social identity theory – twenty-five percent of the sample consisted of minorities who, under social identity theory, would have preferences for their own ingroups. Though running the analyses by restricting the sample to white participants did not change the results, the restricted sample size may have been too small to show any meaningful effects. Thus, to test this theory in the future, the sample should be restricted to white participants.

Another explanation for the mixed support for social identity theory is that whites are the majority racial group in the United States, and ingroup bias increases as a function of the ingroup's proportionate rarity within the population.[185] Further, whites are considered to have high status in the United States,[186] and ingroup bias is greater among low-status groups as compared to high-status groups.[187] As a result, ingroup bias may not be as prevalent among whites as compared with other, lower-status groups. Because whites do not have low status and do not need to increase their status within society, they have less motivation to favor the ingroup.

Further, due to prevailing conceptions of inequality in the criminal justice system,[188] whites may have little motivation to be lenient toward white defendants to combat racism against whites in the justice system and to improve their own self-image, because whites as a group are not stereotyped as criminals. Thus, even when whites have low self-esteem, they may not have enough motivation to express ingroup favoritism to increase self-esteem by further elevating the status of their already high-status ingroup.

In support of this idea, only when individuals perceive that ingroup bias will motivate social change will bias lead to an increase in self-esteem.[189] If one already has high status, there seems to be little need for change. However, it is

185. Brian Mullen et al., *Ingroup Bias as a Function of Salience, Relevance, and Status: An Integration*, 22 EUR. J. SOC. PSYCHOL. 103, 117 (1992).

186. *See* Susan T. Fiske et al., *A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition*, 82 J. PERSONALITY & SOC. PSYCHOL. 878, 892 (2002) (finding that whites are rated highly on both warmth and competence).

187. *See* Mullen et al., *supra* note 185, at 109, 117 (stating that three judges determined status after they read descriptions of the group; high status was when one's ingroup was of higher status than an outgroup; and high-status ingroups exhibit significantly greater biases in the context of artificial groups, but the trend for real high-status groups to exhibit stronger ingroup bias was not significant).

188. *See, e.g.*, DAVID D. COLE, NO EQUAL JUSTICE 9 (1999) (stating that inequality between races in the American criminal justice system is not solely attributable to racism, thereby implying at least part is attributable to racism).

189. Hewstone et al., *supra* note 73, at 580.

still possible that whites think that judging blacks harshly may enact social change; this may be especially true for whites concerned about the closing status gap between blacks and whites.[190] Consequently, it would be important in future studies to measure participants' levels of perceived racial threat to their ingroup to determine whether this affects their propensity to exhibit ingroup bias. Presumably, those low in self-esteem and who perceive a closing of the status gap would be more motivated to discriminate against black defendants to preserve the social status of their ingroup.

2. *Comparison to White Defendant*

In interpreting the data, this Note has consistently equated lower judgments of guilt with lower biases; however, lower judgments of guilt do not necessarily indicate lower biases. In this study, there were only defendants of one race, whereas in the literature, black defendants have almost always been compared to white defendants. This comparison enables researchers to conclude that there are biases against blacks. Without a comparison, it is difficult to know whether the baseline guilt judgments are biased; if they are not biased, the debiasing instructions could actually have created reverse bias and therefore too much leniency toward the black defendant. Despite a lack of comparison, it is likely that biases were reduced in the present experiment, as judgments of guilt differed depending on racism category, suggesting that some participants had biases. In other words, where participants who had high levels of implicit and explicit racism gave higher guilt ratings to a black defendant than participants who had lower levels of implicit and explicit bias, it is reasonable to presume that the high guilt ratings of the racist participants is at least partly attributable to racial bias.

In addition, the choice to have only one race of defendant was deliberate for two reasons. First, another independent variable would have made the study 4 x 2 x 2 x 3 x 2, which is even more unwieldy than what is arguably already a complex study. Second, many ideas tested in the study were new – the implicit and explicit racism levels, debiasing instructions, and timing manipulations. I first wanted to test these variables with a black defendant to see if any of the hypotheses were supported, in the hope of testing any supported variables in a subsequent study comparing black and white defendants. In line with this idea, the next stage of research should pursue a comparison between black and white defendants based on aversive racism theory, as these results mostly supported aversive racism theory. This study could be 2 x 2 x 3 x 2: content of instructions (egalitarian, no instructions), timing (pre-evidence, post-evidence), racism cat-

190. *Id.* at 585.

egory (non-racist, aversive racist, true racist), and race of defendant (black, white). This design would enable the comparison of guilt between black and white defendants and more accurately determine whether egalitarian instructions reduce biases.

I attempted to run such an experiment but found an unusual reverse bias against the white defendant. The origins of this bias are unclear, but I suspect it had to do with the pictures of the white and black defendant used in the experiment. Though taken from the Center for Vital Longevity database[191] and matched for age and gender, there may have been subtle differences between the faces that led participants to think that the white defendant looked more like someone who would commit an assault.[192] One noteworthy conclusion from the experiment is that the egalitarian instructions eliminated this reverse bias.[193] This finding serves to counter the potential objection that a debiasing instruction could cause jurors to overcorrect and favor black defendants, since it appears to make jurors more objective in their judgments regardless of their initial racial preference.

3. *Qualtrics-Based IAT Limitations*

It would be fruitful to employ a new IAT because half of the participants expressed an unusual reverse bias, compared to at most twenty-five percent in the general population,[194] and it would be helpful to see if this was due to the specific measure employed. One such IAT is John Dovidio and colleagues' sub-

191. Minear & Park, *supra* note 126; Park Aging Mind Lab., *supra* note 126.

192. Studies have found that when a suspect's appearance makes the suspect look like someone who would commit a particular crime, such as an assault, that person is judged to be guilty whereas a suspect that does not look like someone who would commit a particular crime is found to be innocent. *E.g.*, C. Neil Macrae & John W. Shepherd, *Do Criminal Stereotypes Mediate Juridic Judgements?*, 28 BRIT. J. SOC. PSYCHOL. 189, 190 (1989).

193. There was a trend interaction between the race of defendant and content of instructions, $F(1, 557) = 2.56$, $p = .110$, $\eta_p^2 = .055$. For the simple effects split by content of instructions, there was a trend in a difference in guilt judgment based on race of defendant when there were no instructions: the white defendant was judged as guiltier than the black defendant, $F(1, 182) = 2.14$, $p = .145$, $\eta_p^2 = .012$. When there were egalitarian instructions, there was no difference in guilt judgment based on race of defendant, suggesting that the egalitarian instructions eliminated the reverse bias against the white defendant.

194. Office of Faculty Dev. & Diversity, *FAQ on Implicit Bias*, STAN. SCH. MED. (2014), http://med.stanford.edu/diversity/FAQ_REDE.html [http://perma.cc/5XNF-7YRS] (reporting that seventy-five percent of Asian and white individuals have an implicit preference for whites over blacks, which means that the remaining twenty-five percent necessarily had no preference or a reverse bias against whites).

liminal IAT.[195] I originally considered using this measure but decided against it because Amazon Mechanical Turk participants would have to download additional software to complete the study. This would be burdensome for participants and would make it more difficult to ensure an adequate sample size, especially if participants encountered computer difficulties with the software. As a result, I decided to use LaBouff's measure. It would be beneficial to replicate my effects with a subliminal IAT, and Dovidio and colleagues' measure would be a good tool for future research.

4. Artificiality of Laboratory Settings

Laboratory studies, though important in investigating hypotheses because of their ability to control confounding variables, have limitations in the context of evaluating jury instructions. These limitations include the following: a laboratory study is much shorter than the length of an actual trial, participants may be less motivated because there is no real defendant whose liberty is at stake, there is no deliberation, and the jury instructions are presented in a written format rather than an oral format. Though these limitations should be considered in interpreting my results, it is important to begin testing a hypothesis through an experiment rather than through a field study.

An experiment enables the researcher to keep all variables constant besides the variables of interest. If I had tested the instructions in actual trials, other variables could have interfered with my results, such as the attractiveness of the defendant, the type of crime, and much more, making it difficult to interpret any results. In addition, it would not have been feasible to test my hypotheses in real trials without any preliminary evidence in support of my hypotheses. It is difficult to imagine that judges would have allowed me to test my instructions in trials, where a defendant faces significant consequences, without any evidence that these instructions are likely to actually reduce biases.

In addition, though the laboratory aspect of my study is a potential limiting factor of its generalizability, many laboratory experiments related to jury instructions have been replicated outside the laboratory. For example, field studies of jury instructions have found that the timing of instructions affects how well jurors can recall evidence,[196] suggesting that timing may not only matter in my study but also in actual trials. In addition, studies of the incom-

195. *See* John F. Dovidio et al., *On the Nature of Prejudice: Automatic and Controlled Processes*, 33 J. EXPERIMENTAL SOC. PSYCHOL. 510 (1997).

196. *See, e.g.*, Heuer & Penrod, *supra* note 108, at 424-25 ("Jurors in criminal trials . . . showed a marginally significant improvement on the [recall] test questions when they were instructed prior to the evidence.").

prehensibility of jury instructions[197] have been replicated in actual trials.[198] Studies such as these suggest that the brevity of laboratory studies, the lack of an actual defendant, the written nature of the instructions, and the lack of deliberation do not necessarily preclude the generalizability of these results to real trials.[199]

CONCLUSION

This Note tested several novel ideas in the race and juror decision-making literature. Rather than observing juror biases and creating post hoc explanations for why such biases occur, this study directly tested two theories against each other. It also examined whether a third race-neutral theory could combat biases. The content of the jury instructions was a novel contribution of the study; this content was derived from principles of aversive racism theory (including the expanded definition proposed in Part II), social identity theory, and procedural justice. In addition, the study measured, for the first time, participants' implicit and explicit racism levels and employed a new method of making race salient to eliminate the confounding factor of interracial crime. The study also manipulated the timing of debiasing instructions to see whether instructions before or after evidence decreased biases to a greater extent.

The data largely support the timing hypothesis and aversive racism theory, but not social identity theory or procedural justice. In line with aversive racism theory, non-racists, aversive racists, and true racists differed in the ratings they gave the black defendant for the variables of guilt judgment, perceived prior record, and sentence judgment, as predicted by their implicit and explicit racism scores. Also consistent with aversive racism theory, the exploratory analysis suggests that egalitarian instructions lower guilt judgments for aversive racists only. Lastly, consistent with previous studies, explicit race salience affected judgments, but only when egalitarian instructions were presented pre-evidence.

Although there are many possible directions for future research, the most intriguing and necessary would focus on aversive racism theory. As previously

197. *See supra* note 115 and accompanying text.

198. *See, e.g.*, Bradley Saxton, *How Well Do Jurors Understand Jury Instructions? A Field Test Using Real Juries and Real Trials in Wyoming*, 33 LAND & WATER L. REV. 59, 109 (1998) (finding that "many of our jurors are misunderstanding at least some of the jury instructions" they are given).

199. In fact, online studies on Mechanical Turk are likely to be even more generalizable than laboratory studies, considering that the sample is more diverse and representative of the general population than the typical undergraduate sample. *See* Paolacci et al., *supra* note 117, at 413.

mentioned, it would be important to determine whether, consistent with aversive racism theory, aversive racists judge black defendants to be guiltier than white defendants when there are no jury instructions, but judge both black and white defendants to be equally guilty when there are egalitarian instructions.

This Note suggests that not only is it important to combat juror biases, but also that the origin of such biases is critical in understanding how to decrease them. The results of this study suggest that guilt judgments are explained more by aversive racism theory than by social identity theory, and that procedural justice-based jury instructions do not decrease biases. Therefore, making race salient in the courtroom and tailoring instructions to egalitarianism are likely to be effective in reducing biases. If, however, social identity theory had explained the results, then a better approach would have been to flatter jurors to increase their self-esteem and to keep race salience low to prevent racial ingroup bias. The differences between these potential applications demonstrate the importance of determining the root cause of juror biases: recommendations for combatting biases can be exactly the opposite depending on which theory explains biases, especially regarding whether to make race salient in a courtroom.

The experimental results from this Note suggest that courts might consider including debiasing elements derived from aversive racism theory in jury instructions. Judges would only need to say a few extra sentences in order to fit the debiasing instructions into current instructions. In addition, presenting these instructions before the evidence phase of a trial would not be difficult and could potentially reduce juror biases in a powerful way.

APPENDIX

Table 1.
DESCRIPTIVE STATISTICS FOR DEPENDENT VARIABLES

Dependent Variables	*Mean*	*SD*
Guilt Judgment	3.50	1.46
Perceived Prior Record	3.55	1.28
Sentence Judgment	3.03	1.39

Note. Ratings were made on a scale of 1 to 7.

Figure 1.
MEAN RATINGS OF GUILT AS A FUNCTION OF RACISM CATEGORY

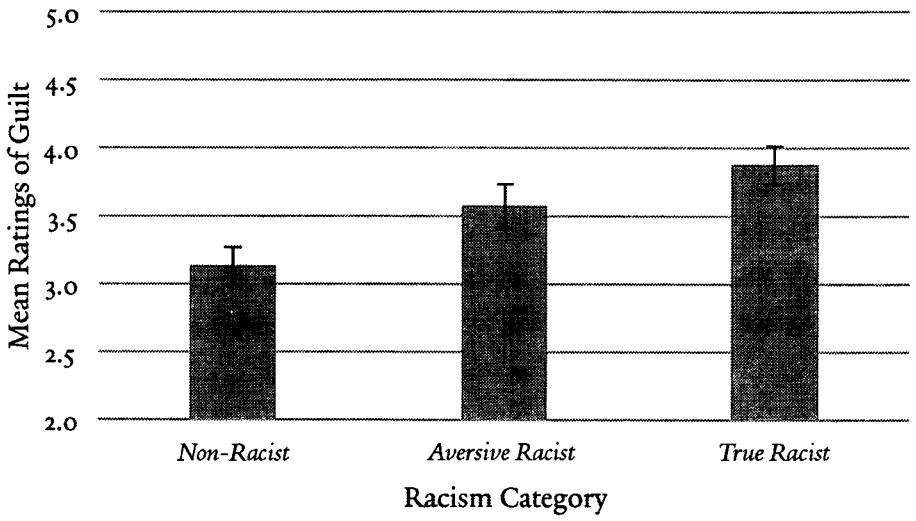

Table 2.
MEAN RATINGS OF GUILT AS A FUNCTION OF RACISM CATEGORY

Racism Category	*Mean Ratings of Guilt*
Non-Racist	3.13_a (.14)
Aversive Racist	3.57_{ab} (.16)
True Racist	3.87_b (.14)

Note. Means that do not share subscripts differ significantly (p = .002) by Tukey's test. Numbers in parentheses refer to standard errors. Guilt ratings were made on a scale of 1 to 7.

Figure 2.
MEAN RATINGS OF GUILT AS A FUNCTION OF CONTENT OF INSTRUCTIONS AND RACISM CATEGORY

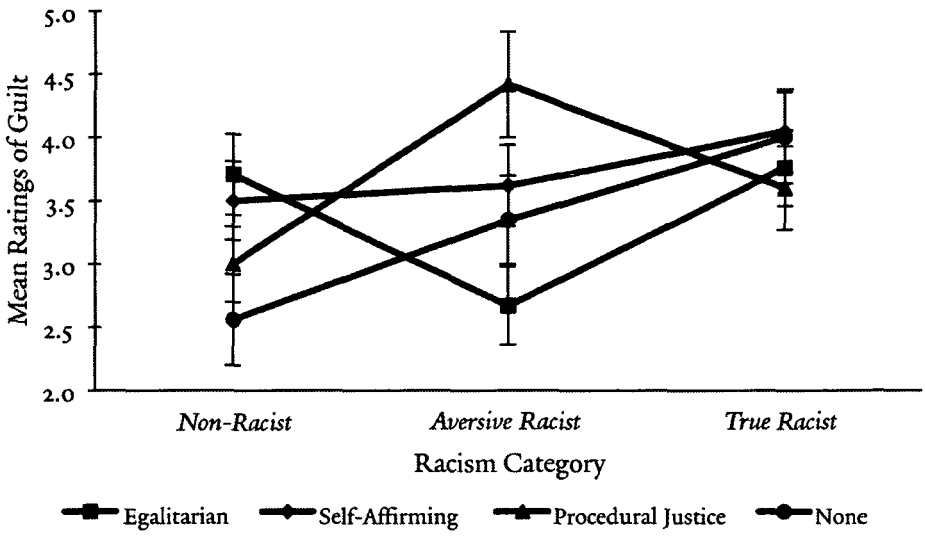

Table 3.
MEAN RATINGS OF GUILT AS A FUNCTION OF CONTENT OF INSTRUCTIONS AND RACISM CATEGORY

	Racism Category		
Instructions	*Non-Racist*	*Aversive Racist*	*True Racist*
Egalitarian	3.71^{a}_{a} (.32)	2.67^{a}_{a} (.44)	3.76^{a}_{a} (.29)
Self-Affirming	3.50^{ab}_{a} (.31)	3.62^{ab}_{a} (.32)	4.05^{a}_{a} (.33)
Procedural Justice	3.00^{ab}_{a} (.30)	4.42^{b}_{b} (.42)	3.60^{a}_{ab} (.33)
None	2.56^{b}_{a} (.36)	3.35^{ab}_{ab} (.35)	4.00^{a}_{b}(.36)

Note. This analysis was exploratory and marginally significant (p = .078). Means that do not share subscripts within each row differ significantly from each other by simple effects. Means that do not share superscripts within each column differ significantly from each other. In the third row (procedural justice) and the first column (non-racist), however, the difference between non-racists and aversive racists and egalitarian and no instructions, respectively, is only marginally significant. Numbers in parentheses refer to standard errors. Guilt ratings were made on a scale of 1 to 7.

Figure 3.
MEAN RATINGS OF PERCEIVED PRIOR RECORD AS A FUNCTION OF RACISM CATEGORY

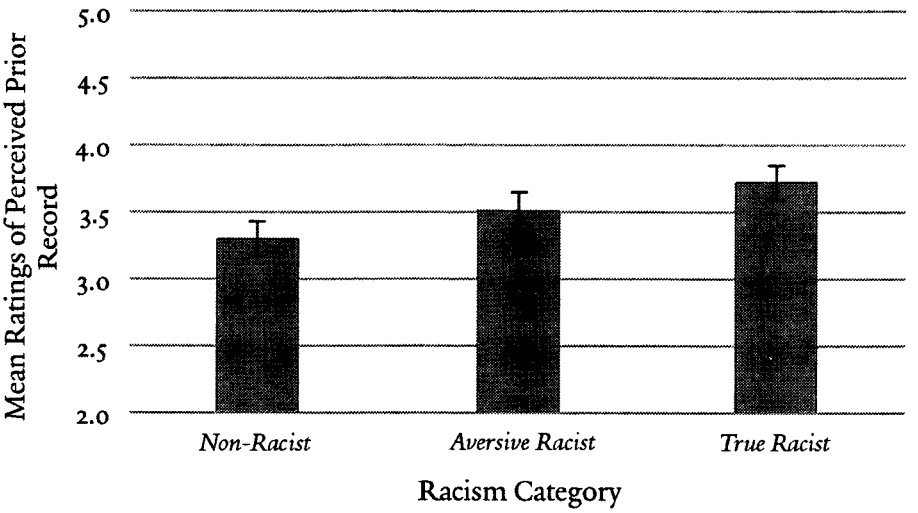

Table 4.
MEAN RATINGS OF PERCEIVED PRIOR RECORD AS A FUNCTION OF RACISM CATEGORY

Racism Category	*Mean Ratings of Perceived Prior Record*
Non-Racist	3.30_{a} (.13)
Aversive Racist	3.51_{ab} (.14)
True Racist	3.72_{b} (.13)

Note. Means that do not share subscripts differ marginally significantly ($p = .067$) by Tukey's test. Numbers in parentheses refer to standard errors. Ratings of perceived prior record were made on a scale of 1 to 7.

Figure 4.
MEAN RATINGS OF SENTENCE JUDGMENT AS A FUNCTION OF RACISM CATEGORY

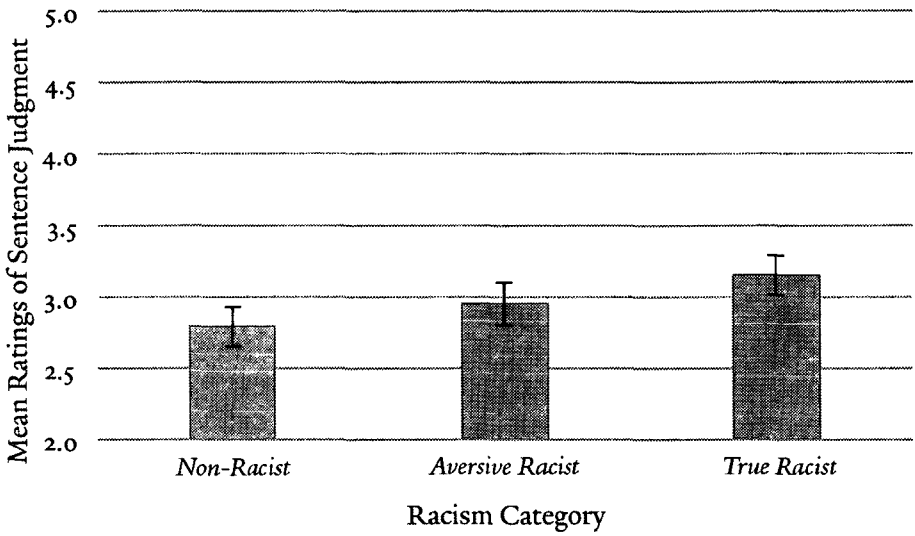

Table 5.
MEAN RATINGS OF SENTENCE JUDGMENT AS A FUNCTION OF RACISM CATEGORY

Racism Category	*Mean Ratings of Sentence Judgment*
Non-Racist	2.79_{a} (.14)
Aversive Racist	2.95_{ab} (.15)
True Racist	3.15_{b} (.14)

Note. Means that do not share subscripts differ marginally significantly ($p = .079$) by Tukey's test. Numbers in parentheses refer to standard errors. Participants indicated the appropriate sentence for the defendant on a scale of 1 to 7.

Figure 5.
MEAN SENTENCE JUDGMENT AS A FUNCTION OF CONTENT OF INSTRUCTIONS, TIMING OF INSTRUCTIONS, AND EXPLICIT SALIENCE

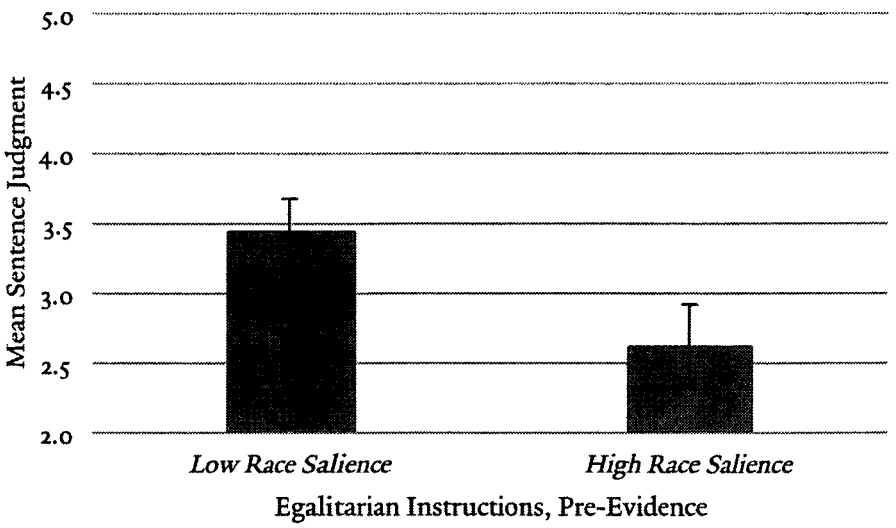

Table 6.
MEAN SENTENCE JUDGMENT AS A FUNCTION OF CONTENT OF INSTRUCTIONS, TIMING OF INSTRUCTIONS, AND EXPLICIT SALIENCE

Content of Instructions	Explicit Salience	Timing of Instructions	
		Pre-Evidence	*Post-Evidence*
Egalitarian	*Low*	3.44_{a} (.24)	2.44_{ab} (.27)
	High	2.62_{b} (.30)	3.24_{ab} (.30)
Self-Affirming	*Low*	3.31_{c} (.27)	2.62_{c} (.38)
	High	3.04_{c} (.27)	2.81_{c} (.22)
Procedural Justice	*Low*	2.86_{d} (.26)	3.75_{d} (.31)
	High	3.05_{d} (.30)	2.96_{d} (.27)
None	*Low*	3.04_{e} (.26)	--
	High	3.24_{e} (.26)	--

Note. Means within each subsection consisting of instructions that do not share any part of a subscript within each row and column differ significantly ($p = .043$) from each other by simple effects. In other words, the only difference is within the egalitarian instructions subsection for the pre-evidence timing of instructions between low and high explicit salience. Numbers in parentheses refer to standard errors. Participants indicated the appropriate sentence for the defendant on a scale of 1 to 7.

Figure 6.
MEAN RATINGS OF GUILT AS A FUNCTION OF TIMING OF INSTRUCTIONS

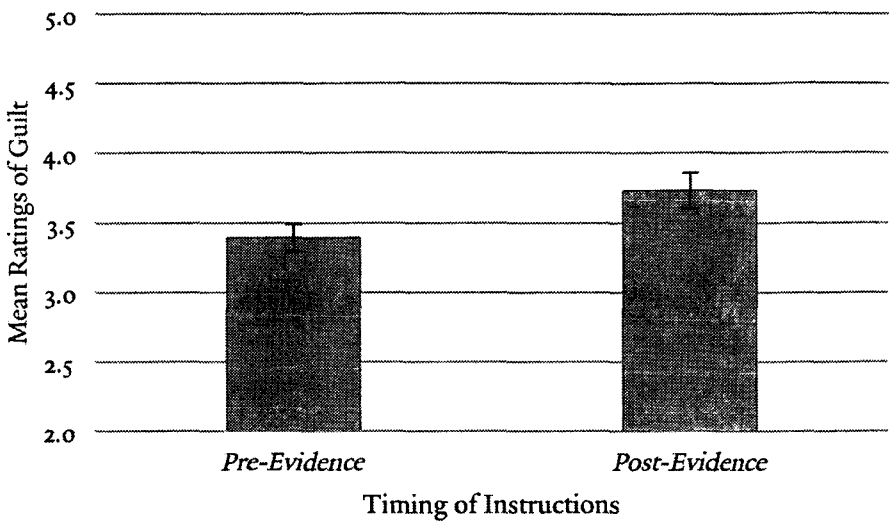

Table 7.
MEAN RATINGS OF GUILT AS A FUNCTION OF TIMING OF INSTRUCTIONS

Timing of Instructions	*Mean Ratings of Guilt*
Pre-Evidence	3.39_a (.10)
Post-Evidence	3.73_b (.13)

Note. Means that do not share subscripts differ marginally significantly ($p = .079$). Numbers in parentheses refer to standard errors.