National Center for State Courts 1971

Can Explicit Instructions Reduce Expressions of Implicit Bias?

New Questions Following a Test of a Specialized Jury Instruction

Jennifer K. Elek & Paula Hannaford-Agor

April 2014

Acknowledgements

The authors would like to thank the State Justice Institute for their generous support of this initiative (SJI-13-N-082; Pilot Test of an Implicit Bias Jury Instruction). In addition, the authors wish to highlight the important contributions of several individuals in this effort. First, we thank Amber Will for her invaluable research and administrative assistance with this study. We also thank our colleagues at the National Center for State Courts for their input and assistance at various stages of this project: Pamela Casey, Diana McSpadden, David Rottman, and Nicole Waters. In addition, we thank our subcontractors, Research Now and Project Implicit, for their technical assistance with the two-part study; in particular, we appreciate the individual attention and support of Amanda Price and Jordan Axt, respectively. Finally, we are extremely grateful for the guidance provided by a consortium of legal and social science scholars with expertise on implicit bias and the law: Devon Carbado, Nilanjana Dasgupta, David Faigman, Phillip Goff, and especially Anthony Greenwald and Jerry Kang. The consortium's involvement in the proposed study is a continuation of an informal scholar-practitioner partnership that emerged from the second annual symposium of PULSE: Program on Understanding Law, Science, and Evidence at the UCLA School of Law in March of 2011.

Abstract

Judges, lawyers, and court staff have long recognized that explicit, or consciously endorsed, racial prejudices have no place in the American justice system. However, more subtle biases or prejudices can operate automatically, without awareness, intent, or conscious control. Members of the court community are beginning to identify this subtler form of racial bias, or implicit racial bias, as a partial explanation for persistent racial disparities in the criminal justice system. In the absence of empirically vetted interventions, some judges have created and currently use their own specialized jury instructions in hopes of minimizing expressions of such bias in juror judgment. However, depending on how these instructions are crafted, they may produce unintended, undesirable effects (e.g., by increasing expressions of bias against socially disadvantaged group members among certain types of individuals, or by making jurors feel more confident about their decision(s) without actually reducing expressions of bias in judgment). To prevent the distribution and implementation of jury instructions that may do more harm than good, any instruction of this kind must be carefully evaluated.

In the present study, the authors sought to examine the efficacy of one specialized implicit bias jury instruction. Mock jurors who received the specialized instruction evaluated the strength of the defense's case in subtly different ways from those who received a control instruction, but the instruction did not appear to significantly influence juror verdict preference, confidence, or sentence severity. Interestingly, the authors were unable to replicate with this sample the traditional baseline pattern of juror bias observed in other similar studies (c.f., Sommers & Ellsworth, 2000; Sommers & Ellsworth, 2001), which prevented a complete test of the value of the instructional intervention. Authors address several possible explanations for this failure to replicate, explore the possibility of shifts in cultural awareness and in the spontaneous correction for bias, and discuss implications for future work.

Introduction

A large body of research evidence indicates that the disparate treatment of racial minorities persists in modern times and pervades all stages of the criminal justice process (e.g., Banks, Eberhardt, & Ross, 2006; The Sentencing Project, 2008; Wooldredge, Griffin, & Rauschenberg, 2005). In the courtroom, even high-stakes verdict and sentencing decisions appear unduly influenced by race bias (e.g., Baldus, Woodworth, & Pulaski, 1990; Mitchell, Haw, Pfeifer, & Meissner, 2005; Rachlinski, Johnson, Wistrich, & Guthrie, 2009). Court leaders across the country have taken aggressive steps to confront this problem over the past three decades (e.g., Casey, Warren, Cheesman, & Elek, 2012; National Center for State Courts, 2007; Spohn, 2000) and in recent years have focused on addressing more subtle or *implicit* forms of racial bias through in-depth education and training of judges and court staff (Casey, Warren, Cheesman, & Elek, 2013; Elek & Hannaford, 2013; Kang, Bennett, Carbado, Casey, Dasgupta, Faigman, Godsil, et al., 2012).

Unlike with judges and court staff, the courts have limited opportunities to educate jurors about the pernicious effects of complex psychological phenomena like implicit bias and how implicit forms of bias may distort jurors' interpretation of trial evidence. Jurors, by definition, are randomly selected to serve from the local community; most jurors in this country serve only for the duration of the trial (typically 2 to 3 days) and then are released from service. There is no time available during this short period to provide the type of in-depth education on strategies to reduce the impact of implicit bias that judges and court staff may receive.

To address the problem of racial bias in juror decision-making, some judges have expressed interest in developing a specialized instruction on implicit bias to include in the set of jury instructions typically proffered on the applicable law governing the case. Judge Mark Bennett (U.S.D.C., Northern District of Iowa) has already created and regularly uses his own implicit bias jury instructions (see Kang et al., 2012). To date, no known studies have examined the effect of a specialized implicit bias jury instruction on expressions of racial bias in jurors' judgments about a case. The authors explore this possibility for the first time in the present study.

Racial Bias in Juror Decision-Making

According to the widely accepted Story Model of juror decision-making (Bennett & Feldman, 1981; Hastie, Penrod, & Pennington, 1983), jurors use the information they receive at trial to construct a narrative or story about the case that is consistent with their world knowledge and that fits the legal categories provided in instructions to the jury. Story construction of this sort is inevitably colored by jurors' personal preconceptions, attitudes, and experiences (i.e., schemas), all of which are used to resolve ambiguities and fill in details missing from evidence and arguments presented at trial. During deliberations, jurors compare elements from their individual narratives (e.g. whether a witness's earlier inconsistent statements means she cannot now be believed) in their effort to arrive at a consensus about the "correct" interpretation of the evidence and the verdict that should follow.

The Story Model approach to understanding juror decision-making helps to explain variation between jurors in individual assessments of the strength of trial evidence. In a 2002 study of hung juries, for example, ratings of the strength of prosecution and defense evidence varied widely between individual jurors serving on the same juries (Hannaford-Agor, Hans, Mott & Munsterman, 2002). Demographic factors accounted for less than three percent of this variation in ratings of evidentiary strength, whereas factors like the perceived importance and credibility of witness testimony played a much larger role. Racial fairness becomes an issue in this context given that racial and ethnic biases have the potential to influence juror perceptions of evidentiary factors (e.g., strength of evidence, eyewitness credibility, attributions of causality) and, through this influence, may inform juror decisions regarding the verdict and sentence (e.g., Levinson, Cai, & Young, 2010; Sommers & Ellsworth, 2000; see Lynch & Haney, 2011).

Overt racial prejudice may explain some of the individual variation in juror judgments, but even individuals who explicitly report egalitarian racial attitudes may nevertheless make racially biased decisions and behave in racially biased ways (see Dasgupta & Asgari, 2004; Dasgupta & Rivera, 2008). This discrepancy may arise in part because (a) explicit self-reports about attitudes are easily contaminated by the respondent's motivated impression management concerns whereas certain behaviors are less easily corrected, and (b) individuals may not be consciously aware of the attitudes and stereotypes they hold that can influence their judgment and behavior (Nosek, 2007). These more subtle cognitions can operate automatically, without awareness, intent, or conscious control, to help shape and to potentially bias decisions (see Greenwald & Banaji, 1995). Although explicit or consciously endorsed racial prejudices in contemporary American society may be on the decline, this "modern," subtler form of racial bias persists (Dovidio, Kawakami, & Gaertner, 2000). Indeed, over the past few decades, various techniques have been used and a number of specialized tests have been developed (such as the popular Implicit Association Test or IAT; Greenwald, McGhee, & Schwartz, 1998) to help researchers identify, measure, and study these so-called *implicit biases* (for a comprehensive review, see Wittenbrink & Schwarz, 2007). These implicit biases are valuable independent predictors of social behavior and judgment in a variety of social and professional decision-making contexts (for a recent meta-analysis, see Greenwald, Poehlman, Uhlmann, & Banaji, 2009; for an review of key studies on the impact of implicit racial bias in settings like voting, hiring, performance assessment, budget setting, policing, and medical treatment, see Jost, Rudman, Blair, Carney, Dasgupta, Glaser, & Hardin, 2009; see also Kang & Lane, 2010; Greenwald & Krieger, 2006).

Minimizing Juror Bias

Social scientists have made great strides in recent years to identify effective (and ineffective) strategies for combating more insidious forms of racial bias. For example, cumulative evidence shows that a multiculturalism approach to egalitarianism (i.e., one that acknowledges group differences and promotes diversity) is more effective in counteracting biases than the popular 'colorblindness' approach that explicitly encourages individuals to ignore race and other differences (e.g., Apfelbaum, Pauker, Sommers, & Ambady, 2010; Apfelbaum, Sommers, & Norton, 2008; Richeson & Nussbaum, 2004; see also Plaut, Thomas, & Goren, 2009). Increased exposure to minority group members who contradict prevailing social stereotypes can also help to reduce implicit racial biases (e.g., Dasgupta & Asgari, 2004; Dasgupta & Greenwald, 2001; Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; for a review of this literature, see Dasgupta, 2009). These and other research findings can inform the development of valuable educational or training programs for judges and other court-employed decision-makers (see also Rudman, Ashmore, & Gary, 2001) in the long term. However, they translate less readily into a viable strategy for use with jurors.

The court's ability to reduce bias in juror decision-making is in many ways restricted; desired solutions are not always feasible. A complete jury trial from start to finish may last only one or two days; time constraints and additional resource limitations (e.g., funding, staffing, technology) often prohibit more elaborate interventions. In this context, viable solutions must already be inherent in the jury trial process or must be easily integrated. Jury deliberations, for example, may help to foster more analytical rather than intuitive or heuristic decision-making, particularly when jurors are prompted to explicitly articulate the basis for their individual verdict preferences (Salerno & Diamond, 2010). When people expect to be held accountable for their decision, they tend to consider a broader array of relevant information, pay more attention to the information they use to support their decision and weight this information more impartially. Through this increased investment of cognitive effort, jurors are more self-aware of the thought process for formulating the decision (see Lerner & Tetlock, 1999). The processing style used depends on the person(s) to whom the decision-maker expects to be held accountable. More racially diverse juries tend to produce decisions less influenced by the defendant's race, presumably for these very reasons (Sommers, 2006). Although more research is needed on the precise mechanisms by which jury diversity affects juror decision-making, it appears that the presence of nonwhites on a jury not only allows for more diverse perspectives to be considered, but may also increase white juror awareness of race-related concerns in a way that stimulates a more thorough and more factually accurate discussion of the evidence. Unfortunately, it is not always possible to ensure such diversity, particularly in jurisdictions with more homogeneous jury pools. Even in more diverse communities, jury panels often fall short of a representative selection of citizens (Sommers, 2008). Thus, other interventions have been proposed.

Historically, courts have relied extensively on jury instructions to guide juror decision-making (Simon, 2012). This approach has been adopted primarily for practical reasons, as instructions are relatively inexpensive, expedient, and easy to administer to each new jury. However, pattern jury instructions developed for use in state and federal jury trials typically rely on the simple admonition that jurors should not let "bias, sympathy, prejudice, or public opinion influence your decision" (Judicial Conference Criminal Jury Instructions, CALCRIM No. 101, 2013). As a next step, judges and lawyers have expressed interest in developing a jury instruction to specifically target the issue of implicit bias.[1]

[1] In addition to the implicit bias jury instructions that Judge Mark Bennett has created and implemented in his own jurisdiction (Kang et al., 2012), an American Bar Association Criminal Justice Section recently assembled a task force to concurrently develop their own instructions (S. Cox and S. Redfield, personal communication, June 3, 2013).

Crafting clear, effective jury instructions on the topic of implicit bias requires extensive subject matter expertise not only to ensure that the language used is an accurate reflection of the state of the science, but also to ensure that the instruction intervention does not incorporate components or delivery approaches known to exacerbate expressions of bias in certain subpopulations. For example, in a direct approach like that adopted by Judge Mark Bennett, instructions could explain the subtle cognitive phenomenon of implicit bias to jurors clearly and in a manner that promotes self-awareness (see Simon, 2012), as individuals can only work to correct for sources of bias that they are aware exist (Wilson & Brekke, 1994) and that they perceive to be self-relevant (see Wegener, Kerr, Fleming, & Petty, 2000). Studies show that individuals can control the behavioral expression of implicit biases in specific laboratory contexts if provided with a concrete strategy for bias reduction (Kim, 2003; Mendoza, Gollwitzer, & Amodio, 2010; Stewart & Payne, 2008). However, strategies which impress an extrinsic motivation to be non-prejudiced (i.e., mandates and other authoritarian language typical of jury instructions) may provoke hostility and resistance, failing to reduce and perhaps even exacerbating expressions of prejudice (e.g., Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Plant & Devine, 2001). Instead, communications designed to foster intrinsic egalitarian motivations may more effectively reduce both explicit and implicit expressions of prejudice (Legault, Gutsell, & Inzlicht, 2011). These and other research findings are important to consider in crafting an effective implicit bias jury instruction.

Any new jury instruction should be carefully evaluated to determine its actual impact on jury decision-making before broadly promoting the instruction as a solution for general use in the courtroom. This is particularly important for the issue of implicit bias, given the possibility that a specialized instruction may successfully reduce expressions of racial bias with some jurors yet exacerbate these expressions (i.e., elicit a backlash effect) among others. To date, no known studies have examined the efficacy of any such jury instruction in reducing racial disparities in juror judgments. For the first time in the present study, authors examined the effect of one specialized implicit bias jury instruction on mock juror judgments.

Method

Design. The present two-part mock trial study used a 2 (defendant race: black or white) x 2 (victim race: black or white) x 2 (instructions: specialized implicit bias or control) factorial design to examine the impact of a specialized implicit bias jury instruction on expressions of racial prejudice in juror decision-making. The two parts of the study were presented to participants as two ostensibly separate and unrelated online surveys, but only participants who completed the first phase of the study (Part 1) were invited to complete the second phase (Part 2).

Participants. To secure a sample of participants, authors contracted with Research Now, an online market research firm with over 6.5 million active panel members. For this experiment, Research Now supplied a nationally representative sample of participants, balanced on age, gender, geographic location, and ethnicity (defined as six possible groupings by Research Now: African American/Black, Asian/Asian American, Caucasian/White, Native American/Inuit/Aleut, Native Hawaiian/Pacific Islander, and Other). To qualify for the present study, authors imposed a set of typical jury eligibility requirements on participation: participants must be U.S. citizens, must reside in the United States, be at least 18 years of age, be able to speak and understand English, and, at the time of the experiment, have no prior felony convictions. Research Now supplied 1287 unique panelists who accessed the experiment online. Of these, 901 panelists (70.0%) completed Part 1 of the study; 386 panelists (30.0%) either failed to meet the minimum eligibility requirements and were thus automatically excluded from further participation or voluntarily chose to withdraw participation by discontinuing with the survey. Of the 901 eligible panelists who completed Part 1, 579 (62.9 %) also completed Part 2 of the study. Of these, 12 participants were excluded based on their IAT error rates (i.e., too many erroneous responses, keystrokes entered too quickly to register as a feasible response to a trial), as recommended by Project Implicit staff (J. Axt, personal communication, August 1, 2013). In addition, due to technical errors in recording the participant identification number, Part 1 and Part 2 responses from six participants could not be matched, leaving 561 participants with a complete, matched set of data from both parts of the study.

Stimulus materials. Authors developed several stimulus materials for use in this experiment. First, four written versions of a mock criminal trial scenario, adapted from Sommers and Ellsworth (2000, 2001), were created in which a defendant was charged with assault and battery with intent to cause serious bodily injury of a victim. In this mock trial scenario, the defendant and victim were described as teammates on a college basketball team and the alleged assault resulted from a locker room altercation.[2] The intent was to utilize a trial scenario with mixed evidence that would elicit a guilty or not guilty verdict in roughly equal proportions**.** Each of the four versions of the mock trial scenario systematically varied the race of the defendant and the race of the victim (see Appendix A).

Second, two versions of jury instructions were constructed and videotaped for use in the experiment. Both versions of the jury instructions were delivered by an older white man in judicial robes from behind a judge's bench. The judge presented standard pattern jury instructions for reasonable doubt (CALCRIM No. 220, 2013), battery causing serious bodily injury (CALCRIM No. 925, 2013) and self-defense (CALCRIM No. 3470, 2013) in both instruction conditions (Appendix B). However, the experimental manipulation focused on the use of a specialized implicit bias jury instruction versus a control instruction of comparable length. Based loosely on a jury instruction developed and used by Judge Mark Bennett of the U.S. District Court, Northern District of Iowa, authors developed a specialized implicit bias jury instruction which incorporates concepts consistent with several promising bias-reduction strategies identified in the broader research literature (for the specialized implicit bias jury instruction with citations to research upon which each component of the instruction is based, see Appendix C). For the control condition instruction, authors selected a pattern instruction of comparable length (adapted from the New York Committee on Criminal Jury Instructions, n.d.), which cautions jurors against conducting internet research about the trial while serving on the jury (Appendix D).

[2] Adaptations from the original Sommers and Ellsworth (2000) scenario included the removal of racially charged language and other cues designed to make race a salient issue to the reader. In addition, the setting of the altercation was described as a college locker room, rather than high school, to avoid special legal considerations and potential biases related to juvenile defendants.

Procedure. Participants completed two ostensibly separate and unrelated web-based surveys in the early summer of 2013. Each survey was estimated to take participants 20 minutes to complete. In exchange for completing each survey, each participant received a "thank you" token reward of through Research Now's online e-Rewards incentive system. The first part of the study (Part 1) connected participants with an experiment designed using Confirmit online survey software. Participants began Part 1 by completing a series of eligibility questions to confirm whether they were indeed jury-eligible. Those who failed to meet the jury eligibility criteria were excluded from further participation in the study.

Participants who met eligibility requirements then received a brief description of the study in which they were asked to assume the role of a juror in a trial case. Participants were randomly assigned to one of eight possible conditions in the experiment: They watched one of the two videotaped sets of jury instructions and then read one of four possible versions of the mock trial scenario describing the evidence in the case against the defendant.

Following the jury instructions and presentation of the evidence in the mock trial, participants then answered a series of questions concerning their verdict preference (guilty/not guilty), confidence in their verdict preference (0% to 100%), assessments of the strength of case for the prosecution and for the defense (7-point Likert scale, 1=extremely weak to 7=extremely strong), and sentencing recommendations (9-point Likert scale, no punishment to maximum punishment of 365 days incarceration). Participants then completed several basic demographic questions to conclude the survey.

Three weeks following the launch of Part 1 and one week following the closure of Part 1 data collection, the second phase of the study (Part 2) launched online. Participants who completed Part 1 of the study were individually invited by Research Now to complete the ostensibly separate Part 2 survey study. To complete Part 2 of the study, respondents were directed to a secure site hosted by Harvard University's Project Implicit, a non-profit research organization with which the authors contracted services. On the Project Implicit hosted survey site, participants completed a measure of race Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT is a popular computer-based measure of implicit attitudes that relies on the individual's response times in a sorting task. This approach is supported by logic that easier pairings (faster responses in the sorting task) reflect stronger associations between concepts (for a complete description, see Greenwald et al., 1998 and Greenwald, Nosek, & Banaji, 2003). In addition, participants answered standard questions provided by Project Implicit that were designed to measure explicit racism by asking how "warm" or "cold" participants felt toward Whites or Blacks, called feeling thermometers. Participants also completed the Symbolic Racism scale (Henry & Sears, 2002) and the Internal and External Motivation to Respond without Prejudice scales (Plant & Devine, 1998).

The explicit and implicit racial bias measures were administered separately from Part 1 to avoid the possibility that either these measures or the mock trial scenario may make race-related norms more accessible to participants than they would otherwise be. Increased salience of race-related norms (c.f. Sommers & Ellsworth, 2006) could potentially alert participants to the primary purpose of the study and influence participant responses on whichever part of the study followed in the sequence if both parts were conducted in a single survey. Authors combined the datasets from Part 1 and Part 2 of the study by matching unique participant identifier numbers supplied by Research Now.

Results

Profile of participants. Of the 901 participants retained for analysis who completed Part 1 of the study, most were female (*n* = 518, 57.5%), white (*n* = 727, 80.7 %), with at least some collegiate-level education (*n* = 803, 89.1%), and of an average age of 50 years.[3] Forty (4.4%) self-identified as Hispanic or Latino/a. The study captured an approximately equal distribution of participants across the West (*n* = 256, 28.4%), Midwest (*n* = 218, 24.2%), South (*n* = 261, 29.0%), and Northeast (*n* = 166, 18.4%) regions of the United States, with the Northeast being slightly underrepresented. Liberal (*n* = 299, 33.2%), moderate (*n* = 246, 27.3%), and conservative (*n* = 356, 39.5%) political attitudes were approximately equally represented. Participant characteristics were distributed across experimental conditions in approximately equal proportions.

Attrition between Part 1 and Part 2 of the study occurred approximately proportionally across all eight experimental conditions and between genders (57.9% female), race and ethnic groups (83.8% white and 3.2% Hispanic), education level (88.7% with at least some college education), geographic region (approximately balanced), political orientation (approximately balanced), and experimental condition (approximately balanced).

Of the 561 participants who completed Part 2 of the study, a large majority exhibited an implicit preference for whites on the race IAT (*n* = 483, 86.1%), with n = 21 (3.7%) exhibiting an implicit preference for blacks and *n* = 57 (10.2%) exhibiting no implicit racial preference (*M* = .61, *SD* = .40).[4] Of the explicit racism feeling thermometer measures, participants on average self-reported slightly warmer feelings towards whites (*M* = 4.17, *SD* = 1.95) than blacks (*M* = 4.76, *SD* = 1.95), with lower scores on each 11-point scale representative of warmer feelings. In addition, to compute the Symbolic Racism 2000 scale index, the eight items on the scale were recoded following procedures used by Henry and Sears (2002) so that 0 = low and 1 = high symbolic racism; items then averaged to create the Symbolic Racism index (α=.831). The scale means for participants in the present study (overall, *M* = 0.46, *SD* = 0.19; for whites, *M* = 0.47, *SD* = 0.19; for nonwhites, *M* = 0.41, *SD* = 0.19) corresponded to means observed by the developers of the Symbolic Racism 2000 scale (Henry & Sears, 2002).

To create the Internal and External Motivation to Respond Without Prejudice scale indices, one item was reverse-coded and participant responses on all items comprising each scale were then averaged to

[3] Of the 19.3% respondents who self-identified as a racial minority, 50 indicated that they were black (5.5%), 77 indicated Asian (8.5%), 7 indicated Native American (0.8%), 5 indicated Hawaiian/Pacific Islander (0.6%), 17 indicated other (1.9%), 16 indicated two or more races (1.8%), and 2 did not provide a selection (0.2%).

[4] The tendency to demonstrate an implicit preference for whites on the race IAT is common in other studies drawing on U.S. samples, particularly among non-black participants (e.g., Nosek, Smyth, Hansen, Devos, Lindner, Ranganath, & Smith, 2007; see also https://implicit.harvard.edu/implicit/demo/background/raceinfo.html).

create the Internal Motivation Scale (IMS) index (α=.813) and the External Motivation Scale (EMS) index (α=.766). Participants were on average externally motivated to respond without prejudice (*M* = 4.90, *SD* = 1.81 on the 9-point EMS index) and highly internally motivated to respond without prejudice (*M* = 7.33, *SD* = 1.64 on the 9-point IMS index); these averages are comparable to results found in the original scale development research (Plant & Devine, 1998).

As expected, several of the racism measures were intercorrelated (Table 1). Of the explicit measures, participants who reported higher symbolic racism also reported cooler feelings towards blacks (*r* = .18, *p* < .01) and warmer feelings towards whites (*r* = -.20, *p* < .01). In addition, participants' IAT scores positively correlated with their explicit Symbolic Racism index scores, *r* = .22, *p* < .01. Participants' IAT raw scores also correlated with their reported feelings towards whites such that stronger implicit preferences for whites were associated with warmer self-reported feelings, *r* = -.13, *p* < .01. Interestingly, participants' IAT raw scores were not significantly related to their self-reported feelings towards blacks, *r* = .05, *ns*. Participants may have engaged in self-monitoring and correction when responding to this particular self-report measure. Participant IAT scores were also positively related to their reported political orientation, with participants who more strongly identified with political conservativism showing a stronger implicit preference for whites, *r* = .12, *p* < .01.

Consistent with prior research (Legault, Green-Demers, Grant, & Chung, 2007), we also found that IMS scores were negatively related to IAT scores, with participants who scored lower on IMS demonstrating a stronger implicit preference for whites, *r* = -.12, *p* <.01. A positive relationship between IAT scores and EMS scores also emerged, with participants who scored higher on EMS demonstrating a stronger implicit preference for whites, *r* = .10, *p* <.05.

Profile of mock trial judgments. Of the 901 participants retained for analysis in Part 1 of the study, 572 (63.5%) submitted a guilty verdict, suggesting that the evidence for defendant culpability was somewhat ambiguous in the mock trial scenario presented to participants. On average, participants reported being fairly confident of their verdict choice (*M* = 77.8% confident, *SD* = 20.53). In addition, participants on average rated the prosecution's case as moderate-to-strong (*M* = 4.87, *SD* = 1.60) and the defense's case as moderate-to-weak (*M* = 3.54, *SD* = 1.63). As expected, perceptions of case strength and verdict choice were intercorrelated: Participant judgments of the strength of the defense's case were negatively related to the reported strength of the prosecution's case (*r* = -.47, *p* < .01) and to the decision to submit a guilty verdict (*r* = -.52, *p* < .01). Similarly, participants' ratings on the strength of the prosecution's case had a positive relationship with judgments of guilt (*r* = .66, *p* < .01). Finally, the most common sentencing recommendation of mock jurors who voted for conviction was for probation only, no incarceration (*n* = 204, or 35.7%), followed by three response options which called for 120 days incarceration or less (*n* = 231, or 40.4%, across all three sentencing options). Only 72 participants (12.6% across three sentencing options) recommended more incarceration time (between 121 and 364 days), and 64 participants selected the maximum penalty of 365 days incarceration (11.2%). One participant recommended no punishment (0.2%).

Participant IAT scores also correlated positively with verdict choice, $r = .09$, $p < .05$. Higher implicit bias scores (i.e., a stronger preference for whites) were associated with the decision to convict the defendant. No other significant relationships were found between participants' IAT score and judgment measures.

Participant EMS scores did not correlate with verdict choice, $r < .01$, *ns*. However, EMS score correlated with other judgment measures. Higher EMS scores were associated with judgments of a stronger case for the prosecution, $r = .68$, $p < .001$, and a weaker case for the defense, $r = -.52$, $p < .001$.

Effects of defendant race on juror judgments. To establish whether or not this experiment successfully replicated previously established juror race bias effects (e.g., Sommers, 2006; Sommers & Ellsworth, 2000, 2001; Cohn, Bucolo, Pride, & Sommers, 2009), we examined if white jurors in control conditions, on average, judged black defendants more harshly than white defendants.[5] Because the race of the victim may not be a critical factor in the expression of juror bias (p. 215, Sommmers & Ellsworth, 2001), we first examined this question across all control conditions, regardless of victim race. We found no evidence that white participants were more likely to convict a black defendant than a white defendant overall, χ^2 (1, $N = 350$) = 1.97, *ns*. Among these participants, we also found no direct effect of defendant race on confidence of verdict, strength of the prosecution's case, strength of the defense's case, or sentence recommendation, *t*s < 1.61. However, an effect of defendant race on the strength of the defense's case approached significance, $t(348)= 1.90$, $p = .06$. Instead of a white juror bias against black defendants, however, white participants in the present study provided higher ratings of the strength of the defense's case when the defendant was described as black ($M = 3.68$, $SD = 1.66$) compared to white ($M = 3.34$, $SD = 1.62$).

To demonstrate the juror bias effect in prior studies, other researchers have commonly relied on interracial trial scenarios as stimulus materials (see Sommers & Ellsworth, 2000, 2001). In a second attempt to establish a juror bias baseline effect, we restricted the analysis to white participants in only those control conditions in which a combination of either a white defendant and black victim or black defendant and white victim was described in the trial scenario.[6] We found no evidence of a white juror bias against black defendants when examining verdict choice in only those control conditions which described an interracial alleged offense, χ^2 (1, $N = 171$) = 0.32, *ns*. We also found no effect of defendant race on confidence of verdict, strength of the prosecution's case, strength of the defense's case, or sentence recommendation, *t*s < 1.59.

[5] Small cell sizes (ns < 5) prohibited a more comprehensive examination of in-group bias or favoritism (e.g., Mitchell, Haw, Pfeifer, & Meissner, 2005; Kerr, Hymes, Anderson, & Weathers, 1995) among other racial or ethnic groups.

[6] Researchers have found stronger expressions of in-group bias or favoritism among Black mock jurors than in White mock jurors in some circumstances (Mitchell et al., 2005; Sommers & Ellsworth, 2009). However, the small subsamples of black mock jurors (*n*s < 5 in some conditions) prohibited further exploration of potential differential effects among racial subgroups.

Effect of instructions. Despite our inability to establish as a baseline the traditional pattern of race bias in the present study, we further explored the effect of the specialized implicit bias jury instruction on participant judgments. We looked first at the entire participant sample, then focused on the judgments of white jurors only,[7] and concluded with an assessment of only those individuals who demonstrated an implicit preference for Whites.

Overall. To examine the effects of the specialized implicit bias jury instructions on participants' reported confidence in their chosen verdict, perceived strength of the prosecution's evidence, perceived strength of the defense evidence , and recommended sentence if found guilty, we conducted 2 (defendant race) x 2 (victim race) x 2 (instruction condition) analysis of variance (ANOVA) tests. A significant three-way interaction effect on strength of the defense's case was observed, F (1, 901) = 3.81, p = .05. No statistically significant effects emerged when the victim was described as white, Fs < 0.96. However, when the victim was described as black, we observed a two-way interaction between instruction condition and defendant race, F(2, 459) = 3.90, p < .05. Among these participants, of those who received the implicit bias instruction, we found no difference in strength of the defense's case when the defendant was described as black (M = 3.56, SD = 1.66) compared to white (M = 3.59, SD = 1.59), F(1, 238) = 0.03, *ns* (Table 2). In the control instruction conditions with black victims, however, participants judged the defense's case to be slightly stronger when the defendant was also described as black (M = 3.85, SD = 1.74) than when the defendant was described as white (M = 3.28, SD = 1.64), F(1, 221) = 6.21, p = .01. Other than a main effect of defendant race explained by this interaction effect, No other statistically significant effects on these variables were observed, Fs < 2.264.

We also found a significant difference in conviction rate by defendant race, χ^2 (1, N = 901) = 4.23, p < .05, with participants convicting the white defendant more often overall (66.8%) than the black defendant (60.2%). No other significant differences were noted, χ^2s < 2.47.

White participants only. We found no differences between conditions in conviction rate among white participants, χ^2s < 2.62. To examine the effects of the specialized implicit bias jury instructions on white participants' reported confidence in their chosen verdict, perceived strength of the prosecution's evidence, perceived strength of the defense evidence , and recommended sentence if found guilty, we conducted 2 (defendant race) x 2 (victim race) x 2 (instruction condition) analysis of variance (ANOVA) tests. A significant three-way interaction effect on strength of the defense's case was observed, F (1, 736) = 6.03, p = .01. No statistically significant effects emerged when the victim was described as white, Fs < 1.45. However, when the victim was described as black, we observed a two-way interaction between instruction condition and defendant race, F(2, 365) = 7.18, p < .01. Among these participants, of those who received the implicit bias instruction, we found no difference in strength of the defense's case when the defendant was described as black (M = 3.47, SD = 1.67) compared to white (M = 3.80, SD = 1.60), F(1, 188) = 1.91, *ns* (Table 3). In the control instruction conditions with black victims, however, participants judged the defense case to be slightly stronger when the defendant was also described as black (M = 3.83, SD = 1.74) than when the defendant was described as white (M = 3.25, SD = 1.59), F(1,

[7] See footnote 6.

177) = 5.392, *p* =.02. In contrast with prior studies that focus on the race of the juror and defendant (Sommers & Ellsworth, 2000, 2001), this pattern underscores the importance of the race of the victim in understanding expressions of juror bias.

Participants with an implicit preference for whites. Among participants who showed an implicit preference for whites on the IAT (n = 483), we found no differences between conditions in conviction rate, χ^2s < 2.61. To examine the effects of the specialized implicit bias jury instructions on these participants' reported confidence in their chosen verdict, perceived strength of the prosecution's evidence, perceived strength of the defense's case, and recommended sentence if found guilty, we conducted 2 (defendant race) x 2 (victim race) x 2 (instruction condition) analysis of variance (ANOVA) tests. No statistically significant effects emerged, *F*s < 3.229.

Probing for a possible backfire effect. To explore for the possibility of a backfire effect among some participants (Plant & Devine, 2001; Legault, Gutsell, & Inzlicht, 2011), we wished to determine whether the specialized implicit bias jury instruction elicited different responses from high- and low-EMS jurors toward black vs. white defendants. We conducted a 2 (instruction condition) x 2 (defendant race) x 2 (juror: primarily IMS vs. primarily EMS) analysis of variance test on strength of the defense's case. We found no evidence of a backfire effect, *F*(1, 526) = 0.11, *ns*. However, because so few jurors in this study were primarily externally motivated to respond without prejudice (n = 76), these results should be interpreted with caution. Small cell sizes prohibited further exploration.

Discussion

The present study found no significant effects of the instruction on judgments of guilt, confidence, strength of the prosecution's evidence, or sentence length. Unexpectedly, the control conditions of the present study failed to generate the traditional patterns of juror bias, in which white mock jurors judge black defendants more harshly than white defendants. Without replicating this pattern of bias to establish a baseline against which participants in the experimental conditions could be compared, we were unable to fully examine the effectiveness of the specialized implicit bias jury instruction in reducing bias in juror judgments.

Despite our inability to replicate the traditional juror bias effect in this study, we uncovered some evidence to suggest that the specialized implicit bias jury instruction could influence juror appraisals in a mock trial case. Participants who received these specialized instructions prior to reading the trial scenario produced a different pattern of judgments of the strength of the defense's case compared with participants who received a control instruction. Specifically, in control conditions, jurors indicated that the defense's case was stronger (in fact, the strongest of all eight conditions) when the alleged offense occurred between a black defendant and a black victim, rather than between a white defendant and a black victim. However, the specialized implicit bias jury instructions tempered this racial disparity. Further research could explore why the black defendant – black victim crime produced the highest strength-of-case ratings for the defense. Was it perceived as easier to defend, or as most easily justified? Perhaps the present study failed to detect the traditional pattern of juror bias because people are getting better at detecting straightforward racial issues and correcting for potential bias in their judgments about a defendant, but not yet good at detecting or correcting for more complex expressions of racial bias.

In addition, we found no evidence to suggest that the specialized instruction produces a harmful backfire effect among those likely to feel threatened by and react against external pressure to comply with mandatory non-discrimination standards (Plant & Devine, 2001; Legault, Gutsell, & Inzlicht, 2011). These findings should be considered strictly preliminary. Future research efforts should oversample individuals who are more externally motivated to respond without prejudice to permit a more conclusive analysis and to further probe for possible differential effects. For an instruction approach to remain a feasible solution in the context of jury trials, judges must be able to administer the instruction to the jury as a whole without risking a backfire effect among an unknown proportion of the trial's jurors.

Possible explanations for the failure to replicate the traditional juror bias effect. Ultimately, we found no conclusive evidence regarding whether or not the specialized implicit bias jury instruction used in this study is effective as a bias-reduction intervention, primarily because the present study failed to replicate the original baseline effect of juror bias documented by Sommers and Ellsworth (2001) using the original version of the same stimulus materials. Without a replication of the baseline pattern of juror bias, a clear test of the full value of the specialized implicit bias jury instruction is not possible. Many potential explanations for these findings focus on differences between the present study and the original Sommers and Ellsworth (2000) study, such as:

(1) Differences in materials. For example, the slight modifications we made to the original Sommers and Ellsworth (2000) stimulus materials effectively eliminated the expected effects. Alternatively, perhaps the traditional pattern jury instructions on reasonable doubt, battery causing serious bodily injury, and self-defense used in the present study were sufficient to wipe out the juror bias effect observed in Sommers and Ellsworth (2000), making the specialized instruction superfluous.

(2) Differences in setting. In the present study, participants completed surveys individually online. However, Sommers and Ellsworth (2000) conducted the original study in person, in group settings. Moreover, their participants were relatively homogeneous assemblies of fraternity and sorority members on a single college campus who likely knew one another prior to the experiment. It is possible that in the original study, the presence of others from the participants' own social in-group could have influenced participant judgment in ways that differed meaningfully from the present study, and in ways that may not generalize to the typical jury.[8] Alternatively, the responses from the current sample of jury-eligible adult web users may be

[8] For an example of how even racial homogeneity of otherwise unfamiliar fellow jury members can influence juror decision-making, see Sommers (2006). See also Sommers (2007) for further discussion regarding the influence of the composition of a jury on the juror decision-making process.

unique because of factors in the uncontrolled environment in which the experiment was completed.

(3) Differences in the composition of participant samples. Sommers & Ellsworth (2000) observed the original juror bias effect with a sample of college students, whereas the present study used a national sample of jury-eligible adult web users. It is possible that findings among samples of college students do not generalize well to other samples, such as the one used in this experiment (see, generally, Wiener, Krauss, & Liberman, 2011).

However, the above explanations for the failure to replicate Sommers and Ellsworth's (2001) original juror bias finding using nearly identical stimulus materials are unlikely. The juror bias effect documented by Sommers and Ellsworth (2000) has been replicated in group and individual participant settings, with college students and broader community samples of jury-eligible adults, using different trial scenarios as stimulus materials, and using methods to present those materials with more and less ecological validity, such as by asking mock jurors to read a short trial summary, watch a videotaped summary presentation, or participate in a large-scale trial simulation (e.g., Sommers & Ellsworth, 2000; Sommers & Ellsworth, 2001; Cohn, Bucolo, Pride, & Sommers, 2009). Moreover, the failure in the current experiment to replicate the Sommers and Ellsworth (2000) baseline pattern of juror bias has since been duplicated: At least three other research studies conducted after the present experiment (between November 2013 and January 2014) also failed to replicate the original juror bias effect, with college students and with two samples of web users recruited through Amazon Mechanical Turk (P. Ellsworth, personal communication, February 16, 2014). This provides convergent evidence of what is potentially a broader contemporaneous shift in the pattern of bias expression.

Although one might argue from these findings that the problem of implicit bias has been "solved," such a conclusion is both premature and unlikely. Even in the present experiment, the majority of participants exhibited a strong implicit race bias in favor of Whites. Indeed, other research continues to demonstrate the persistence of implicit forms of bias in general and as expressed in social judgment and behavior (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Jost, Rudman, Blair, Carney, Dasgupta, Glaser, & Hardin, 2009; Kang & Lane, 2010). A more likely explanation for the failure to replicate the original Sommers and Ellsworth (2000) juror bias effect is that Americans may have become increasingly aware of the cultural attention to race bias over the past decade and are now more sensitive to the possibility of revealing such bias, particularly in research settings. This heightened level of awareness and sensitivity may trigger spontaneous self-correction of the kind measured (and in some cases, experimentally manipulated) in prior research (e.g., see Green, et al., 2007; Sommers & Ellsworth, 2000; Sommers & Ellsworth, 2001). Whether this sensitivity is temporary or a more permanent reflection of the modern age is another empirical question to be answered: The trends observed in the present study could be the first sign of a permanent change in the efforts of Americans to self-monitor and correct for expressions of bias, but it is possible that contemporaneous media attention on race bias and the justice system throughout 2013, such as the Zimmerman and Alexander "Stand Your Ground" trials in Florida, could have served to temporarily enhance awareness of and sensitivity to these issues. The increased salience of race and race norms in routine media communications about the American justice system could have primed participants to spontaneously self-monitor and correct for possible bias. If the latter is the case, a simple reminder to consider race and race norms may be sufficient to prompt jurors to engage in corrective action against expressions of bias in judgment.

Until some of the above questions are addressed, the verdict on any question regarding the utility of specialized implicit bias jury instructions cannot be rendered. Although highly unlikely, if everyone is routinely engaging in self-monitoring and correction of bias in decision-making, there is no need to engage in future efforts to develop and test specialized implicit bias jury instructions. However, if some individuals do not spontaneously engage in self-correction or if people inconsistently engage in self-correction, further research is needed to determine whether or not the technique in general can effectively and more consistently trigger corrective action. If the technique works generally, additional research could identify the instructional components that most effectively trigger corrective action for inclusion in a recommended model instruction. If warranted, future research should account for more complex effects: A specialized implicit bias jury instruction may produce desired bias-reduction effects on only particular subpopulations of individuals or under particular conditions. Jury deliberations could amplify the intended effect (and/or backfire effect) of a well-crafted implicit bias jury instruction, and meaningful effects may be observed only when providing the instruction to actual jurors in real world trials with real consequences. Alternatively, it remains possible that a specialized implicit bias jury instruction (including or limited to the specific instruction developed for and tested in the present study) simply will not work as a bias-reduction intervention.

Beyond a specialized implicit bias jury instruction, researchers could explore the utility of any of several other justice system intervention strategies for reducing the effects of implicit bias on judgment (e.g., see Casey, Warren, Cheesman, & Elek, 2012; Elek & Hannaford, 2013). These strategies show promise but have not yet received sufficient empirical scrutiny. Because of the potential for a backfire effect and the possibility of “doing harm,” these approaches should be fully vetted by research scientists before recommending practical implementation.

References

Aarts, H., Gollwitzer, P., & Hassin, R. (2004). Goal contagion: Perceiving is for pursuing. *Journal of Personality and Social Psychology, 87*, 23-37.

Alter, A., & Oppenheimer, D. (2009). Suppressing secrecy through metacognitive ease: Cognitive fluency encourages self-disclosure. *Psychological Science, 20*, 1414-1420.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In blind pursuit of racial equality? *Psychological Science, 21,* 1587-1592. doi: 10.1177/0956797610384741

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, *95*, 918-932.

Baldus, D., Woodworth, G., & Pulaski, C. (1990). *Equal justice and the death penalty: A legal and empirical analysis.* Boston: Northeastern University Press.

Banks, R., Eberhardt, J., & Ross, L. (2006). Discrimination and implicit bias in a racially unequal society. *California Law Review, 94,* 1169-1190.

Bennett, W. & Feldman, M., (1981). *Reconstructing reality in the courtroom justice and judgment in American culture.* New Brunswick, NJ: Rutgers University Press.

Casey, P., Warren, R., Cheesman, F., & Elek, J. (2012). *Helping courts address implicit bias: Resources for education*. Williamsburg, VA: National Center for State Courts.

Casey, P., Warren, R., Cheesman, F., Elek, J. (2013). Addressing implicit bias in the courts. *Court Review, 49*, 64-70.

Clark, R., & Maass, A. (1988). The role of social categorization and perceived source credibility in minority influence. *European Journal of Social Psychology, 18*, 381-394.

Cohn, E., Bucolo, D., Pride, M., & Sommers, S. (2009). Reducing white juror bias: The role of race salience and racial attitudes. *Journal of Applied Social Psychology, 39*, 1953-1973. doi: 10.1111/j.1559-1816.2009.00511.x

Dasgupta, N. (2009). Mechanisms underlying the malleability of implicit prejudice and stereotypes: The role of automaticity and cognitive control. In T. Nelson (Ed). *Handbook of prejudice, stereotyping, and discrimination* (pp. 267-284). New York: Psychology Press.

Dasgupta, N. & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40*, 642-658.

Dasgupta, N., & Greenwald, A. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800-814.

Dasgupta, N., & Rivera, L. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition, 26*, 54-66.

Devine, P., Plant, E., Amodio, D., Harmon-Jones, E., & Vance, S. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835-848.

Djikic, M., Langer, E., & Stapleton, S. (2008). Reducing stereotyping through mindfulness. *Journal of Adult Development, 15*, 106-111.

Dovidio, J., Kawakami, K., & Gaertner, S. (2000). Reducing contemporary prejudice: Combating explicit and implicit bias at the individual and intergroup level. In S. Oskamp (Ed.), *Reducing prejudice and discrimination: The Claremont Symposium on applied social psychology* (pp. 137-163). Mahwah, NJ: Lawrence Erlbaum Associates.

Elek, J.K., & Hannaford-Agor, P. (2013). First, do no harm: On addressing the problem of implicit bias in juror decision-making. *Court Review, 49*, 190-198.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and ingroup favoritism. *Journal of Personality and Social Psychology, 78*, 708-724.

Green, R., Carney, D., Pallin, D., Ngo, L., Raymond, K., Iezzoni, L., & Banaji, M. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine, 22*, 1231-1238.

Greenwald, A., & Banaji, M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4-27.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*, 945-967.

Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Greenwald, A., Nosek, B., & Banaji, M. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*,197-216.

Greenwald, A., Poehlman, T., Uhlmann, E., & Banaji, M. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17-41. doi: 10.1037/a0015575

Guthrie, C., Rachlinski, J., & Wistrich, A. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review, 93,* 106-109.

Hannaford, P.L., Hans, V.P., Mott, N.L., & Munsterman, G.T. (2002). *Are hung juries a problem?* Williamsburg, VA: National Center for State Courts.

Hastie, R., Penrod, S. & Pennington, N. (1983). *Inside the Jury Room.* Cambridge: Harvard University Press.

Henry, P.J. and Sears, D. O. (2002), The Symbolic Racism 2000 scale. *Political Psychology, 23,* 253–283. doi: 10.1111/0162-895X.00281

Jost, J., Rudman, L., Blair, I., Carney, D., Dasgupta, N., Glaser, J., et al. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29,* 39-69.

Judicial Council of California Criminal Jury Instructions (2013). *CALCRIM No. 101, Cautionary admonitions: Jury conduct.* San Francisco: LexisNexis.

Judicial Council of California Criminal Jury Instructions (2013). *CALCRIM No. 220, Reasonable doubt.* San Francisco: LexisNexis.

Judicial Council of California Criminal Jury Instructions (2013). *CALCRIM No. 925, Battery causing serious bodily injury.* San Francisco: LexisNexis.

Judicial Council of California Criminal Jury Instructions (2013). *CALCRIM No. 3470, Self-defense.* San Francisco: LexisNexis.

Kang, J., Bennett, M., Carbado, D., Casey, P., Dasgupta, N., et al. (2012). Implicit bias in the courtroom. *UCLA Law Review, 59*, 1124-1186.

Kang, J., Dasgupta, N., Yogeeswaran, K., & Blasi, G. (2010). Are ideal litigators white? Measuring the myth of colorblindness. *Journal of Empirical Legal Studies, 7,* 886-915.

Kang, J., & Lane, K. (2010). Seeing through colorblindness: Implicit bias and the law. *UCLA Law Review, 58,* 465-520.

Kawakami, K., Dovidio, J., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871-888.

Kim, D. (2003) Voluntary controllability of the implicit association test (IAT). *Social Psychology Quarterly, 66*, 83-96.

Langer, E., Bashner, R., & Chanowitz, B. (1985). Decreasing prejudice by increasing discrimination. *Journal of Personality and Social Psychology, 49*, 113-120.

Legault, L., Green-Demers, I., Grant, P., & Chung, J. (2007). On the self-regulation of implicit and explicit prejudice: A Self-Determination Theory perspective. *Personality and Social Psychology Bulletin, 33*, 732-749.

Legault, L., Gutsell, J., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22*, 1472-1477.

Levinson, J., Cai, H., & Young, D. (2010). Guilty by implicit bias: The guilty-not guilty implicit association test. *Ohio State Journal of Criminal Law, 8*, 187-208.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Lynch, M. & Haney, C. (2011). Looking across the empathic divide: racialized decision making on the capital jury. *Michigan State Law Review, 2011*, 573-607.

Mendoza, S., Gollwitzer, P., & Amodio, D. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin, 36*, 512-523.

Mitchell, T., Haw, R., Pfeifer, J., & Meissner, C. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior, 29*, 621-637.

National Center for State Courts (2007). *Interactive database of state programs.* Retrieved from http://www.ncsconline.org/D_Research/ref/programs.asp

N.Y. Committee on Criminal Jury Instructions (n.d.). *Required jury admonitions*. Retrieved from http://www.nycourts.gov/judges/cji/5-SampleCharges/CJI2d.Preliminary_Instructions.pdf

Nosek, B. (2007). Implicit-explicit relations. *Current Directions in Psychological Science, 16*, 65-69.

Pfeifer, J., & Ogloff, J. (1991). Ambiguity and guilt determinations: A modern racism perspective. *Journal of Applied Psychology*, *21*, 1721 – 1725. doi: 10.1111/j.1559-1816.1991.tb00500.x

Plant, E., & Devine, P. (1998). Internal and external motivation to respond without prejudice. *Journal of Psychology and Social Psychology*, *75*, 811-832.

Plant, E. & Devine, P. (2001). Responses to other-imposed pro-black pressure: Acceptance or backlash? *Journal of Experimental Social Psychology, 37*, 486-501. doi: 10.1006/jesp.2001.1478

Plaut, V., Thomas, K., & Goren, M. (2009). Is multiculturalism or color blindness better for minorities? *Psychological Science, 20,* 444-446.

Rachlinski, J., Johnson, S., Wistrich, A., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84*, 1195-1246.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338-342.

Richeson, J., & Nussbaum, R. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology, 40*, 417-423.

Rudman, L., Ashmore, R., & Gary, M. (2001). “Unlearning” automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology, 81*, 856-868.

Rose, M., Diamond, S., Butler, K. (2010). Goffman on the jury: Real jurors’ attention to the “offstage” of trials. *Law & Human Behavior, 34*, 310-323.

Salerno, J.M. & Diamond, S.S. (2010). The promise of a cognitive perspective on jury deliberation. *Psychonomic Bulletin Review, 17*, 174-179.

Sechrist, G., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. *Journal of Personality and Social Psychology, 80*, 645-654.

The Sentencing Project (2008). *Reducing racial disparity in the criminal justice system: A manual for practitioners and policymakers.* Washington, DC: Authors.

Simon, D. (2012). More problems with criminal trials: The limited effectiveness of legal mechanisms. *Duke Law Journal, 75,* 167-209.

Sommers, S. (2006). On racial diversity and group decision-making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology, 90*, 597-612.

Sommers, S. (2008). Determinants and consequences of jury racial diversity: Empirical findings, implications, and directions for future research. *Social Issues and Policy Review, 2*, 65-102

Sommers, S., & Ellsworth, P. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin, 26,* 1367-1379.

Sommers, S., & Ellsworth, P. (2001). White juror bias: An investigation of prejudice against Black defendants in the American courtroom. *Psychology, Public Policy, and Law, 7,* 201-229.

Son Hing, L., Li, W., & Zanna, M. (2002). Inducing hypocrisy to reduce prejudicial responses among aversive racists. *Journal of Experimental Social Psychology, 38*, 71-78.

Spohn, C.C. (2000). Thirty years of sentencing reform: The quest for a racially neutral sentencing process. In J. Horney (Ed.), *Criminal justice 2000: Vol. 3. Policies, processes, and decisions of the criminal justice system* (pp. 427–501). Washington, DC: U.S. Department of Justice, National Institute of Justice.

Stewart, B., & Payne, B. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34*, 1332-1335.

Wegener, D., Kerr, N., Fleming, M., & Petty, R. (2000). Flexible corrections of juror judgments: Implications for jury instructions. *Psychology, Public Policy, and Law, 6,* 629-654.

Wegener, D., & Petty, R. (1995). Flexible correction processes in social judgment: The role of naïve theories in corrections for perceived bias. *Journal of Personality and Social Psychology, 68*, 36-51.

Wegener, D., & Petty, R. (1997). The flexible correction model: The role of naïve theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology, 29*, 141-208.

Wiener, R., Krauss, D., & Lieberman, J., Eds. (2011). Special issue: Jury simulation research: Student versus community samples. *Behavioral Sciences and the Law, 29,* 325-479.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116*, 117-142.

Wittenbrink, B., & Schwarz, N., Eds. (2007). *Implicit measures of attitudes*. New York: Guilford.

Wooldredge, J., Griffin, T., & Rauschenberg, F. (2005). (Un)anticipated effects of sentencing reform on the disparate treatment of defendants. *Law & Society Review, 39,* 835-873.

Yuki, M., Maddux, W., Brewer, M., & Takemura, K., (2005). Cross-cultural differences in relationship- and group-based trust. *Personality and Social Psychology Bulletin, 31*, 48-62.

Table 1. Correlations Between Implicit and Explicit Measures of Racial Bias (n = 561)

	IAT raw score	Symbolic Racism Index	Black Thermometer	White Thermometer	EMS Index	IMS index
IAT raw score	-					
Symbolic Racism Index	.22**	-				
Black Thermometer	.05	.18**	-			
White Thermometer	-.13**	-.20**	.63**	-		
EMS Index	.10*	.08	-.03	-.19**	-	
IMS Index	-.12**	.42**	-.37**	-.02	.04	-

**p* < .05, ** *p* < .01

Table 2. Mean Ratings for Strength of the Defense's Case as a Function of Instruction Condition, Defendant Race, and Victim Race Among All Jurors (n = 901)

	Control Instruction				Implicit Bias Instruction			
	White Defendant		Black Defendant		White Defendant		Black Defendant	
	M	SD	M	SD	M	SD	M	SD
White Victim	3.48	(1.62)	3.54	(1.56)	3.35	(1.57)	3.66	(1.64)
Black Victim	3.28	(1.64)	3.85	(1.74)	3.59	(1.59)	3.56	(1.66)

Table 3. Mean Ratings for Strength of the Defense's Case as a Function of Instruction Condition, Defendant Race, and Victim Race Among White Jurors (n = 736)

	Control Instruction				Implicit Bias Instruction			
	White Defendant		Black Defendant		White Defendant		Black Defendant	
	M	SD	M	SD	M	SD	M	SD
White Victim	3.44	(1.64)	3.51	(1.56)	3.31	(1.52)	3.64	(1.62)
Black Victim	3.25	(1.59)	3.83	(1.74)	3.80	(1.60)	3.47	(1.67)

Appendix A

Mock Trial Scenario

***State v. Matthew Stevenson,* Superior Court no. CR12-563425.**

Defendant: Matthew Stevenson, 21-year-old [white or black] male, 6’ 4” 210 lbs.

Victim: Rod Bentley, 19-year-old [white or black] male, 6’ 2” 192 lbs.

The prosecution charges that the defendant, Matthew Stevenson, is guilty of battery with serious bodily injury. Stevenson was the starting point guard on the basketball team for Johnson State College, but the team had been struggling, and the coach decided to bench him in favor of Rod Bentley, a younger, less experienced player. Before the next game, Stevenson approached Bentley in the locker room and began yelling at him. Witnesses explain that the frustrated defendant told Bentley that he was a “fuckin’ bench warmer” and he couldn’t wait to put him “back in his place.” When another teammate, Jacob Thompson, stepped between the two players, Stevenson shoved him and told him to get out of the way. The prosecution claims that Bentley then grabbed Stevenson to separate him from Thompson, but the defendant threw Bentley off, pushed him into a row of lockers, and ran out of the room. As a result of this fall, two of Bentley’s teeth were broken and he was knocked unconscious. Bentley now suffers from a permanent 80% loss of hearing in his right ear as a result of this assault. The prosecution claims that Stevenson has shown no remorse for his crime and has even expressed to friends that Bentley “only got what he had coming.”

The defense claims that Stevenson was merely acting in self-defense and that Bentley’s injuries were accidental. Stevenson felt he had been the subject of nasty remarks and unfair criticism throughout the season from his teammates. Stevenson claims that he was afraid during the altercation in the locker room. He admits he “might have said something inappropriate to Bentley,” but he says that he was just frustrated and it was nothing worse than what he had heard from the rest of the team all season. Stevenson claims that when Bentley then grabbed him, he felt that he was in danger and tried to break free, and that he must have accidentally knocked into Bentley in the attempt to get out of the locker room. He explained that the reason he never apologized to Bentley in the hospital was that he knew no one on the team would’ve visited him if he’d been the one hurt, but he did say that it was “a shame” that Bentley had been injured so seriously.

Appendix B

Standard Pattern Jury Instructions

Reasonable Doubt

The fact that a criminal charge has been filed against the defendant, [defendant name], is not evidence that the charge is true. You must not be biased against Mr. Stevenson just because he has been arrested, charged with a crime, or brought to trial.

A defendant in a criminal case is presumed to be innocent. This presumption requires that the prosecution prove a defendant guilty beyond a reasonable doubt. Whenever I tell you the prosecution must prove something, I mean they must prove it beyond a reasonable doubt.

Proof beyond a reasonable doubt is proof that leaves you with an abiding conviction that the charge is true. The evidence need not eliminate all possible doubt because everything in life is open to some possible or imaginary doubt.

In deciding whether the prosecution has proved its case beyond a reasonable doubt, you must impartially compare and consider all the evidence that was received throughout the entire trial. Unless the evidence proves [defendant name] guilty beyond a reasonable doubt, he is entitled to an acquittal and you must find him not guilty.

Battery Causing Serious Bodily Injury

[Defendant name] is charged with battery causing serious bodily injury. To prove that he is guilty of this charge, the prosecution must prove that:

1. [Defendant name] willfully touched the victim, [victim name], in a harmful or offensive manner; AND
2. [Victim name] suffered serious bodily injury as a result of the force used; AND
3. [Defendant name] did not act in self-defense.

Someone commits an act *willfully* when he or she does it willingly or on purpose. It is not required that he or she intend to break the law, hurt someone else, or gain any advantage.

Making contact with another person, including through his or her clothing, is enough to commit a battery.

Self-Defense

Self-defense is a defense to battery. [Defendant name] is not guilty of that crime if he used force against another person in lawful self-defense. [Defendant name] acted in lawful self-defense if:

1. He reasonably believed that he was in imminent danger of suffering bodily injury; AND

2. He reasonably believed that the immediate use of force was necessary to defend against that danger; AND

3. He used no more force than was reasonably necessary to defend against that danger.

When deciding whether [defendant name]'s beliefs were reasonable, consider all the circumstances as they were known to and appeared to him and consider what a reasonable person in a similar situation with similar knowledge would have believed. If [defendant name]'s beliefs were reasonable, the danger does not need to have actually existed.

Appendix C

Annotated Specialized Implicit Bias Jury Instruction

Our system of justice depends on the willingness and ability of judges like me and jurors like you to make careful and fair decisions.[9] What we are asked to do is difficult because of a universal challenge: We all have biases. We each make assumptions and have our own stereotypes, prejudices, and fears.[10] These biases can influence how we categorize the information we take in.[11] They can influence the evidence we see and hear, and how we perceive a person or a situation. They can affect the evidence we remember and how we remember it. And they can influence the "gut feelings" and conclusions we form about people and events.[12] When we are aware of these biases, we can at least try to fight them.[13] But we are often not aware that they exist.

We can only correct for hidden biases when we recognize them and how they affect us. For this reason, you are encouraged to thoroughly and carefully examine your decision-making process to ensure that the conclusions you draw are a fair reflection of the law and the evidence.[14] Please examine your reasoning for possible bias by reconsidering your first impressions of the people and evidence in this case. Is it easier to believe statements or evidence when presented by people who are more like you?[15] If you or the people involved in this case were from different backgrounds – richer or poorer, more or less educated, older or younger, or of a different gender, race, religion, or sexual orientation – would you still view them, and the evidence, the same way?[16]

Please also listen to the other jurors during deliberations, who may be from different backgrounds and who will be viewing this case in light of their own insights, assumptions, and

[9] When leadership sets an egalitarian example, others may also pursue this goal (see Aarts, Gollwitzer, & Hassin, 2004).

[10] To avoid potential backfire effects, instructional language should reduce external pressure to comply (by avoiding authoritarian language) and promote intrinsic motivation to counteract biases (Plant & Devine, 2001; Richeson & Nussbaum, 2004; Legault, Gutsell, & Inzlicht, 2011).

[11] See Guthrie, Rachlinski, and Wistrich (2007); Legault et al. (2011)

[12] See Levinson, Cai, and Young (2010); Plant and Devine (1998); Kang, Dasgupta, Yogeeswaran, and Blasi (2010). On the effects of bias on perception and judgment, and regarding how awareness of potential bias may help trigger self-correction efforts, see Green, Carney, Pallin, Ngo, Raymond, Iezzoni, and Banaji (2007).

[13] People often are not aware of their own biases. For people to attempt to correct for bias, they must know that it is a problem and also believe the problem to be self-relevant (see Wilson & Brekke, 1994; see also Wegener, Kerr, Fleming, & Petty, 2000; Wegener & Petty, 1995; 1997).

[14] A more deliberative mode of thinking may help to reduce expressions of bias (see Langer, Bashner, & Chanowitz, 1985; Djikic, Langer, & Stapleton, 2008; see also see Guthrie et al., 2007; Pfeifer & Ogloff, 1991).

[15] For a discussion on processing fluency and perceptions of trust, see Reber and Schwarz (1999); Alter and Oppenheimer (2009). See also Clark and Maass (1988), Yuki, Maddux, Brewer, and Takemura (2005).

[16] Perspective-taking may help to reduce the accessibility and expression of stereotypes (see Galinsky & Moskowitz, 2000).

even biases.[17] Listening to different perspectives may help you to better identify the possible effects these hidden biases may have on decision-making.[18]

Our system of justice relies on each of us to contribute toward a fair and informed verdict in this case. Working together, we can reach a fair result.[19]

[17] Instructions which encourage people to attend to and appreciate one another's differences (i.e., a multiculturalism philosophy) are more effective at reducing expressions of bias than instructions to ignore individual differences (i.e., a colorblindness philosophy); the latter may induce a backfire effect, thereby increasing expressions of prejudice (e.g., Apfelbaum, Sommers, & Norton, 2008). Note that the mere presence of a racial minority on a panel of mostly white jurors may reduce the likelihood of a biased verdict (Sommers, 2006).

[18] See Wegener et al. (2000). When individuals are held accountable for the decision-making process itself, they tend to think more deliberatively; however, when they are only held accountable for the outcome, they may be more likely to attempt to justify unjust decisions retrospectively (see Lerner & Tetlock, 1999). These instructions are designed to focus the juror on the process. In addition, if people are made aware of their biases, those who endorse egalitarianism but remain implicitly biased may actively correct for bias in their decision-making (see Son Hing, Li, & Zanna, 2002). If in the presence of a relatively egalitarian-minded group, an individual's judgments may become less stereotypic (see Sechrist & Stangor, 2001).

[19] Emphasizes goal-setting and leadership involvement; see footnote 9.

Appendix D

Control Condition Instruction

Juror Conduct During the Trial

Do not converse, either among yourselves or with anyone else, about anything related to the case.

Do not visit the place where the crime was allegedly committed. You must not use Internet maps, or Google Earth, or any other program or device to search for any location discussed in the testimony.

Do not read or listen to any accounts of the case reported by newspapers, television, radio, the Internet, or any other news media.

Do not attempt to research any fact or law related to this case, whether by discussion with others or by research on the Internet.

I want to emphasize that you must not communicate with anyone about the case by any means, including telephone, text messages, email, Internet chat or chat rooms, blogs, or social Web sites such as Facebook, MySpace, or Twitter.

You must not provide any information about the case to anyone by any means, including posting information about the case, or what you are doing in the case, on any device or Internet site. You also must not Google or otherwise search for any information about the case, or the law which applies to the case, or the people involved in the case, including the defendant, the witnesses, the lawyers, or the judge.