Manuscript under review (2002)

Bias in the absence of malice: The paradox of unintentional deliberative discrimination

Robert W. Livingston

University of Wisconsin—Madison

Abstract

The literature on attitude-behavior correspondence generally maintains that deliberative (as opposed to spontaneous) discrimination is controllable, related to explicit attitudes, and regulated by motivation to control prejudice. Three studies test the conditionality of each of these findings by investigating the nature of deliberative discrimination against ethnic minorities when response ambiguity is high. Results indicate that deliberative discrimination was positively related to implicit as well as explicit racial attitudes. Moreover, racial attitudes predicted discrimination even when controlling for motivation to control prejudice, whereas the relationship between motivation to control prejudice and discrimination disappeared when controlling for prejudice. Finally, results reveal that discriminatory bias persisted in the face of strong manipulations of accountability, fear of validity, and the presence of a minority experimenter. As a whole, these results provide strong evidence that even highly deliberative discrimination can occur independent of intent or motivation, and suggest that theoretical conceptualizations of "control" incorporate epistemic factors as well as factors of time and agency.

"All animals are equal, but some are more equal than others"

---George Orwell, *Animal Farm*

Social inequality in some form or another plagues every society on earth. The insidious effects of differential treatment can be seen across a variety of spheres including education, healthcare, and the legal justice system (Sidanius & Pratto, 1999). However, the antecedents of institutional discrimination remain nebulous, particularly when it emerges in societies that strive to uphold egalitarian ideals. According to many social psychologists, the occurrence of discriminatory bias among individuals who have truly internalized egalitarian values depends on the extent to which the behavior in question is controllable. Current dual-process models classify discriminatory responses into two broad categories that vary along a continuum of control, with spontaneous or automatic responses occurring independent of awareness or egalitarian intent, and deliberative or controlled responses being regulated by individual differences in motivation to control prejudice (Devine, 1989; Devine & Monteith, 1999; Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Dovidio, Kawakami, & Gaertner, 2002; Fazio, 1990; Fazio & Towles-Schwen, 1999; Wilson, Lindsey, & Schooler, 2000).

According to Fazio's MODE model, for example, spontaneous discrimination (e.g., nonverbal behaviors) is determined by automatically activated racial attitudes, independent of motivation to control prejudice, because the "opportunity" (i.e., time, control, or cognitive resources) to respond in manner consistent with one's motivation is low. However, when opportunity is high, such as the case with deliberative discrimination, motivation to control prejudice rather than racial attitudes determines the occurrence of discriminatory outcomes. Thus, a high prejudiced individual who is high in motivation to control prejudice would be

expected show evidence of discriminatory bias on a spontaneous task, whereas the same individual would not show evidence of discriminatory bias on a deliberative task, because the latter affords sufficient opportunity to override the influence of racially prejudiced attitudes.

In contrast to this position, other theories have argued that attitudes can exert unconscious or unintended influences on even deliberative behavior, independent of motivation (Wilson & Brekke, 1994). The term "mental contamination" coined by Wilson and Brekke (1994) refers to the "process whereby a person has an unwanted judgment, emotion, or behavior because of mental processing that is unconscious or uncontrollable. By unwanted, we mean that the person making the judgment would prefer not to be influenced in the way he or she was" (Wilson & Brekke, 1994; p.117). Individuals are typically unaware of the impact of preexisting attitudes on deliberative decisions and behaviors. For example, "when teachers assign a C to a student's paper, they probably believe that they have given it a fair and unbiased evaluation, even if they were biased by how much they like the student" (Wilson & Brekke, 1994; p. 121).[1]

In a similar vein, the aversive racism model argues that "good intentions are not sufficient to guarantee that equal opportunity will ensure equal treatment" (Dovidio & Gaertner, 1996; p. 67). This model further posits that unintentional racial bias is most likely to occur in situations in which a clear, nonbiased response is not readily apparent, or when individuals can attribute potentially biased responses to some factor other than race (Dovidio & Gaertner, 1996; Gaertner & Dovidio, 1986). Situations involving response or attributional ambiguity are more likely to produce unintentional bias because individuals are unsure what a nonbiased response entails, and are effectively unable to respond in a way that is consistent with their egalitarian values (albeit for lack of knowledge rather than lack of control per se). Devine and colleagues also argue that prejudiced behaviors may derive from a lack of knowledge of how to respond in

intergroup settings rather than a lack of motivation to behave in a nonprejudiced fashion (see Devine & Vasquez, 1998).

Despite theoretical support, there is no direct empirical support for the paradoxical notion that deliberative discrimination can occur independent of motivation to control prejudice. The present study sought to provide such a test. The logic here is that if deliberative discrimination is the product of intent, then motivation to control prejudice should override racial attitudes in determining behavioral outcomes (Fazio, 1990). However, if discrimination occurs unintentionally as the result of mental contamination, then attitudes should predict behavioral outcomes independent of motivation (Wilson & Brekke, 1994). There were no strong predictions as to whether implicit or explicit attitudes, or both, would predict discrimination. On the one hand, research has generally found that implicit attitudes tend to correlate with spontaneous behaviors, while explicit prejudice correlates with deliberative behaviors (Dovidio et al., 1997; Dovidio, Kawakami, & Gaertner, 2002; Fazio, Jackson, Dunton, & Williams, 1995; Wilson, Lindsey, & Schooler, 2000). However, other research has shown that implicit attitudes can crossover to predict deliberative behaviors (Swanson, Rudman, & Greenwald, 2001; Vargas, von Hippel, & Petty, 2002), particularly under conditions of ambiguity (Sargent & Theil, 2002).

Experiment 1

Method

Participants. Sixty-eight White students (24 males, 44 females) participated in partial fulfillment of course requirements.

Procedure. Experiment 1 was divided into two sections: (1) a primary session in which students completed explicit measures of racial attitudes, the five-item internal motivation to respond without prejudice scale (IMS; Plant & Devine, 1999)[2], and the criminal sentencing task,

and (2) a follow-up session one to three weeks later in which students completed the measure of implicit racial attitudes. Explicit racial attitudes were assessed with a packet of feeling thermometers that measured participants' attitudes toward a variety of groups including "Hispanics (Latinos)", and "White Americans (Caucasians)". Ratings were given on a 101-point scale with 0 labeled as "very cold" and 100 labeled as "very warm". Ratings of Hispanics were subtracted from ratings of Whites to create a prejudice index. Higher numbers indicate more anti-Hispanic prejudice.

Implicit racial attitudes were measured using a Hispanic version of the Implicit Association Task (IAT; Greenwald, McGhee, & Schwartz, 1999). The basic design of the IAT involves judgments of two categories of words: Ethnicity (Hispanic or White) and Evaluation (pleasant or unpleasant). During the IAT participants are presented with words that are either Hispanic names (e.g., Juan) or Anglo names (e.g., John), and their task is to decide whether the name is Hispanic or White by pressing one of two computer keys labeled "Hispanic" or "White". Interspersed with these name presentations are presentations of words that are either positive (e.g., flower) or negative (e.g., death) in connotation. Their task for these words will be to judge wither the word is "pleasant" or "unpleasant" by pressing a key labeling accordingly. Reaction times to judge all words are recorded in milliseconds. The hallmark of the task is that the category labels are associated with the evaluative label, such that the "Hispanic" and "unpleasant" key are one and the same, and the "White" and "pleasant" key are one in the same, or vice versa. The participants then complete two blocks of word presentations in which "Hispanic" and "unpleasant" are paired, for example, and the other in which "Hispanic" and "pleasant" are paired. The underlying assumption in this task is that participants high in implicit prejudice against Hispanics should take longer to respond to words when "Hispanic" and

"pleasant" are paired, compared to reaction times to respond to words when "Hispanic" and "unpleasant" are paired (see Greenwald, McGhee, & Schwarz, 1997, for full description).

Deliberative discrimination was operationalized in terms of a criminal sentencing paradigm, in which there was no one "correct" response, but rather a wide range of plausible response options. Nearly all of the psychological and legal research on prejudice and discrimination has focused on African Americans (e.g., Devine, 1989; Dovidio & Gaertner, 1996; Fazio, Jackson, Dunton, & Williams, 1995; Gaertner & Dovidio, 1986; McConahay, 1986; Sommers & Ellsworth, 2000), so it is unclear whether many of these theoretical and empirical findings extend to other ethnic groups for whom historical factors, socialization patterns, or social norms may differ. Consequently, Study 1 used Hispanics, now the most populous and fastest growing ethnic group in the United States, as the target ethnic group.

Participants were randomly assigned to ethnicity conditions and given two descriptions of crimes allegedly obtained from the police department of a large midwestern city. The first crime was a filler scenario that dealt with a university student who had committed an act of cruelty against a domesticated animal. The second crime, which served as the critical crime scenario, described either a White or Hispanic male who assaulted a White female. Defendant ethnicity was manipulated using both name and photograph. Participants were informed that all defendants had pled guilty, so their only job was to decide on the appropriate sentence. Participants were not under any time pressure to read the crime scenario or make sentencing decisions.

The specific crime description for the White condition is listed below. Items in bold in parentheses indicate modifications to the paragraph in the Hispanic condition.

The assault in question took place in the [city district], at approximately 10:30 p.m. on November 15, 1998. The perpetrator of the crime, **David Edmonds (Juan Luis Martinez)**

moved to [city] from **Canada (Mexico)** 10 years ago to take a job with the [city] Country Club. The victim, 28 year old, Carol Wilkins was walking with her friend Nancy Balderston when the incident took place. Witnesses say that **David (Juan)**, who was slightly intoxicated at the time, began yelling foul and distasteful comments at Carol on the night of November 15th. The two ladies decided to ignore **David (Juan)**, which apparently made him angrier, at which point he approached the ladies and began behaving aggressively. When Carol told Juan to go away and leave them alone, he became hostile and began to physically assault Carol. Nancy then ran into a nearby bar to ask for help and call the police. When she returned, she found that Carol had been badly injured and suffered a head concussion, some fractured ribs, a broken nose, and required over a dozen stitches. **David (Juan)** had left the scene by the time the police arrived, but was arrested at his home two days later. **David (Juan)** was charged with aggravated assault and battery. This was **David's** (**Juan's**) third criminal offense.

Because prior research has shown that bias is less likely to occur under conditions in which race is salient (Sommers & Ellsworth, 2000), half of the participants were randomly assigned to receive instructions that included a paragraph that warned participants of the potential for racial bias and urged them to avoid such bias. Instructions to the other half of the participants made no mention of race or instructions to avoid bias. The race salience/motivation paragraph appeared in bold and read as follows:

You should make every effort to be fair when sentencing. Prior research has shown that sentencing decisions can be unintentionally influenced by extraneous factors such as the race…of the offender. Please try not to let these factors influence you decisions.

For the sentencing phase, participants selected one of fourteen levels of punishment within six distinct categories. The categories of punishment and levels of punishment (in parentheses) within each category were: (1) verbal reprimand, (2-6) probation, (7-16) prison sentence, (17) disenfranchisement and deportation, (18) the death penalty. Only probation and prison sentences contained different levels. The specific levels for the probation option were: (2) 30 days, (3) 60 days, (4) 90 days, (5) 6 months, (6) 1 year. For prison, they were: (7) 7 days, (8) 30 days, (9) 90 days, (10) 6 months, (11) 12 months, (12) 3 years, (13) 5 years, (14) 7 years, (15) 10 years, or (16) 25+ years. This sentencing measure provided an 18-point scale of sentencing

severity with (1) verbal reprimand as the most lenient and (18) death penalty as the most severe. Participants were then thanked, thoroughly debriefed, and dismissed.

Results

To ascertain the existence of any discriminatory bias, as well as whether such bias differed as a function of race salience/motivation manipulation, a 2 (Ethnicity: Mexican vs. White) x 2 (Salience/Motivation Instructions: Absent vs. Present) ANOVA was performed with sentence severity as the dependent variable. Results revealed a significant main effect of ethnicity, $F(1, 64) = 4.24$, $p = .04$, such that Hispanic targets received harsher sentences than White targets ($M = 13.31$ and $M = 11.94$, respectively). However, the salience/motivation manipulation had no main effect on discrimination, $F(1, 64) = .01$, $p < .92$, nor did it interact with defendant ethnicity, $F(1, 64) = .17$, $p < .68$ (see Figure 1).

The next set of analyses examined the relationship between discrimination and individual differences in prejudice or motivation to control prejudice. As seen in Table 1, there was only a marginally significant correlation between implicit prejudice and discrimination, $r = .37$, $p < .10$, indicating that implicit bias against Hispanics was associated with harsher punishment for Hispanic defendants. None of the other correlations approached significance.

Discussion

Study 1 obtained significant evidence of discrimination against Hispanic defendants compared to White defendants for the exact same offense. Moreover, this bias was unmoderated by manipulations designed to raise the salience of race and motivation to avoid bias.[3] Further, discriminatory bias was also unrelated to individual differences in motivation to control prejudice. There was, however, a marginally significant relationship between discrimination and implicit (but not explicit) prejudice, suggesting that participants with automatic cognitive

associations between Hispanics and negativity assigned harsher punishments to Hispanic defendants vis-à-vis White defendants.

Although these results are consistent with the notion that discrimination occurs unintentionally, they only partially support the mental contamination hypothesis, and raise some methodological concerns. For instance, the null relationship between discrimination and motivation to control prejudice could have occurred because the specific measure used was inappropriate for Latino targets. As previously mentioned, the IMS was designed specifically for Blacks, so it is unclear whether the modified version validly taps motivation to control anti-Hispanic prejudice. Also, the null relationship between discrimination and explicit prejudice could have been due to problems with the racial attitude measure. There are few if any good scales of Hispanic racial attitudes and the feeling thermometer is a only a single item measure. In addition, the stronger relationship between discrimination and implicit measures may have occurred not because of level of implicitness per se, but because the IAT, unlike the feeling thermometer, assesses evaluations of the outgroup in repeated trials, or because the IAT was administered in a separate session weeks later.

Alternatively, the weak correlations obtained in Experiment 1 may have resulted from factors related to the target group rather than the measures per se. The feeling thermometer, even being a single item measure, has proven to be effective in a multitude of previous studies across the social sciences. Moreover, although the IMS was not originally designed for use with Hispanics, manipulated motivation to control prejudice also failed to moderate discrimination. Thus, perhaps the weak correlations for Latinos was due to relatively weak attitudes and/or weak social norms proscribing prejudice against this group. Prior research has shown that stronger or more accessible attitudes are more likely to predict behavior compared with weaker attitudes

(Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Krosnick & Petty, 1995), and that motivation to control prejudice works better for groups for whom norms against prejudice are strong (Crandall, Eshleman, & O'Brien, 2002). Given the considerable differences in intergroup contact, socialization patterns, and historical conflict with Blacks compared with Latinos in the region where this study was conducted, it is likely that attitudes toward African Americans are much stronger than those toward Latinos. Prior research has also shown that social norms against anti-Black prejudice tend to be stronger than social norms against anti-Hispanic prejudice (Crandall et al., 2002) To test the possibility that the particular ethnic group membership of the target may moderate the relationship between discrimination and attitudinal and motivational variables, Experiment 2 included both Black and Latino targets.

Experiment 2

Participants. Ninety-four students participated in partial fulfillment of course requirements. The data from three Hispanic and three Black participants were omitted. The final sample consisted of 88 participants (21 males, 67 females).

Method. The method of this experiment was essentially similar to that of Experiment 1, with the exception of an additional level of the ethnicity factor. Participants were randomly assigned to read about a White, Hispanic, or Black defendant. Feeling thermometers were added to assess attitudes against "African Americans (Blacks)" in addition to "Hispanics (Latinos)", and "Whites (Caucasians)". Those in the Black defendant condition completed a black version of the IAT which included African American names (e.g., Malik, Aisha) in lieu of Hispanic names. Finally, participants completed the IAT immediately after completing to the sentencing task, in the same session, as opposed to completing the IAT in a separate session weeks later.

Results

Consistent with Experiment 1, Hispanic defendants received harsher sentences than did White defendants for the identical criminal offense, $t(55) = 2.28$, $p < .03$ (Ms = 12.74 and 11.26, respectively). Moreover, black defendants also received harsher sentences compared with White defendants, $t (52) = 2.14$, $p < .04$ (Ms = 12.81 and 11.26, respectively), while there was no difference in sentencing severity between Hispanic and Black targets, $t (63) = .10$, $p = .92$. Next, correlations between discriminatory bias and implicit prejudice, explicit prejudice, or motivation to control prejudice were assessed separately for White, Hispanic, and Black targets and are displayed in Table 2.[4]

Consistent with Experiment 1, correlations were generally weak for Hispanic targets, all $ps > .17$ (see Table 2). However, for Black targets, there was a significant positive correlation between explicit prejudice and discrimination, $r = .70$, $p < .005$, as well as implicit prejudice and discrimination, $r = .43$, $p < .015$, indicating that those higher in prejudice showed greater discriminatory bias against Black defendants. Additionally, there was a marginally significant negative correlation between motivation to control prejudice and discrimination, $r = -.318$, $p < .081$, such that lower levels of motivation to control prejudice were associated with higher levels of discrimination.

To test whether this pattern of relationships actually differentiated responses toward Black vis-à-vis White defendants, three separate simultaneous regression analyses were computed in which the dependent variable was sentence severity, and the independent variables were: (1) ethnicity (with White coded as 0 and Black coded as 1), (2) implicit prejudice, explicit prejudice, or IMS (centered), and (3) an ethnicity x variable interaction term. As shown in Table 3, significant interactions emerged between race and both explicit prejudice and motivation to control prejudice, beta = .44, $p < .02$ and beta = -.51, $p < .04$, respectively. Additionally, a

marginally significant race x implicit prejudice interaction emerged, beta = .48, $p < .08$. The same analyses performed for Latino vis-à-vis White defendants found only a marginally significant effect of motivation to control prejudice, beta = -.40 $p < .06$ (see Table 3).

Finally, given the significant relationship between explicit prejudice and motivation to control prejudice, $r = -.60$, $p < .023$, and implicit prejudice and motivation to control prejudice, $r = -.58$, $p < .001$ for participants in the Black condition (the relationship between implicit and explicit prejudice was nonsignificant, $r = .29$, $p = .31$), the last set of analyses focused on the critical question of whether discriminatory bias against Blacks was mediated by racial attitudes or motivation to control prejudice. Two regression analyses were computed separately for implicit prejudice and explicit prejudice. The dependent variable in both cases was sentence severity, which was regressed on (1) defendant ethnicity, (2) motivation to control prejudice, (3) prejudice (implicit or explicit), (4) ethnicity x prejudice (implicit or explicit), and (5) ethnicity x motivation. Results are presented in Table 4. Looking at the equation for explicit prejudice, $F (5, 23) = 5.83$, $p < .001$, results show that only the explicit prejudice x ethnicity interaction term obtained significance, beta = .60, $p < .006$. The same equation computed for implicit prejudice, $F (5, 48) = 3.07$, $p < .02$, yielded no significant predictors (see Table 4).

Discussion

To date, most evidence for the mental contamination of behavior has been theoretical or anecdotal in nature. There have yet to be any direct empirical evidence that attitudes predict deliberative responses independent of motivation to control prejudice. However, the present study provide compelling evidence that racial attitudes predict deliberative discrimination independent of motivation to control prejudice, despite the strong relationship between motivation and explicit prejudice. Further, the significant negative relationship between

motivation and discrimination disappeared when prejudice was added to the regression equation.[5] In short, although a substantial portion of what is expressed in explicit attitudes reflects motivation to avoid prejudice, these attitudes continued to predict deliberative discrimination even when this shared variance was partialled out.

Although the magnitude of discriminatory bias against Black and Hispanic targets was nearly identical, the underlying correlates were quite different, suggesting that current models of prejudice focusing on Blacks (e.g., McConahay, 1986; Gaertner & Dovidio, 1986; Fazio et al., 1995) may not be generalizable to other ethnic minority groups. As seen in Table 2 and Table 3, racial attitudes predicted discriminatory bias against Blacks only, suggesting that the attitude-behavior correspondence was moderated by attitude strength. Further, discriminatory bias against Blacks was predicted by implicit as well as explicit measures of prejudice.[6] As previously stated, some models have generally maintained that explicit attitudes influence deliberative, well-considered responses whereas implicit attitudes affect responses that are more difficult to monitor or control, such as nonverbal behaviors. The present findings demonstrate that a strict interpretation of this implicit-spontaneous, explicit-deliberative formulation may be too rigid. As Experiment 2 demonstrates, as well as other research (Sargent & Theil, 2002; Vargas, von Hippel, & Petty, 2002), implicit attitudes may sometimes predict deliberative behaviors. Even Dovidio and colleagues argue that while explicit measures tend to be more strongly related to deliberative discrimination, implicit measures tend to be more related to spontaneous actions, this is not a strict dichotomy. Even some of their studies have obtained marginal evidence of crossover effects (e.g., Dovidio et al., 1997; Experiment 3). The results of Experiment 2, as well as those of other research (i.e., Sargent & Theil, 2002), indicate that such a crossover is more likely to emerge under conditions of response ambiguity.

While Experiment 1 indicated that manipulations designed to raise awareness and motivation were unsuccessful in attenuating bias, and Experiment 2 showed that the moderation of deliberative discrimination by motivation was accounted for by shared variance with racial attitudes, it seemed worthwhile to test once more for any effects of motivation on discrimination. Experiment 3 manipulates rather than measures motivation. However, the manipulations designed to raise motivation and eliminate bias in Experiment 3 were more numerous and much stronger than those used in Experiment 1, introducing elements of accountability (Lerner & Tetlock, 1999), fear of invalidity (Kruglanski & Freund, 1983), and the influence of a minority experimenter (Fazio et al., 1995).

Experiment 3

Method

Participants. Seventy-seven White college students from introductory psychology courses participated in partial fulfillment of course requirements.

Procedure. The basic design and procedure of Experiment 3 was similar to that of Experiment 1 with the addition of the bias reduction manipulations that are described below. For starters, components of accountability were introduced. While participants in prior studies believed that their responses were anonymous, participants in Study 2 were instructed to write their full names and e-mail addresses on the front of the packet, thereby increasing identifiability (see Lerner & Tetlock, 1999). Secondly, participants were also told that they would have to reveal and justify their sentencing decisions to the experimenter and to other participants in the study, thereby increasing reason-giving (Lerner & Tetlock, 1999). In addition, fear of invalidity was induced by informing participants that the study involved a panel of circuit judges interested in civilian perceptions of crime and punishment. They were informed that the judges would be reviewing

their juridical decisions and might contact them at some point in the future to discuss their sentence recommendations. Finally, because past research has shown that the expression of discriminatory bias against minorities can be affected by the presence of minorities (e.g., Fazio et al.,1995), Experiment 3 was conducted by a minority experimenter. In short, Experiment 3 contained a barrage of strong inducements for participants to avoid discriminatory bias if possible. After completing the experiment, students were then thanked, debriefed and dismissed.

Results

To assess whether the discriminatory bias persisted in the face of strong bias-reduction manipulations, a one-way ANOVA was performed with defendant ethnicity as the independent variable and sentence severity as the dependent variable. This analysis revealed a marginally significant effect of ethnicity, $F(1, 75) = 3.37$, $p = .07$, indicating once again that the Hispanic defendant received harsher sentences than the White defendant ($M = 9.42$ and $M = 8.22$, respectively). While the manipulations did not eliminate bias, they did seem to considerably lower overall sentence severity. Although there was no control condition in this study, Figure 2 presents the pattern of sentencing across the three studies.[7] As seen in Figure 2, the manipulations in Study 3 had a sizeable impact on the overall sentencing of defendants compared with Experiments 1 and 2. While mean sentence severity was 12.63 in Experiment 1, and 12.00 in Experiment 2, the mean sentencing severity was only 8.82 in Experiment 3. Treating Experiment as a categorical variable, a 3 (Experiment: 1, 2, or 3) x 2 (Ethnicity: Latino or White) ANOVA revealed, consistent with the pattern of data displayed in Figure 2, a highly significant effect of Experiment, $F(2, 196) = 41.89$, $p < .0001$. Post hoc analyses reveal that there is was a significant difference in overall sentence severity between Study 1 and Study 3, $p < .0001$, and Study 2 and Study 3, $p < .0001$, but no significant different between Study 1 and Study 2, $p <$

.68. Moreover, there was a highly significant meta-analytic effect of ethnicity, $F(1, 196) = 12.36$, $p < .001$, affirming the robustness of discriminatory bias. However, there was absolutely no evidence of an interaction between Experiment and Ethnicity, $F(2, 196) = .05$, $p < .95$, suggesting that the bias-reduction manipulations did not significantly attenuate bias against Latino defendants in Experiment 3 compared to Experiments 1 or 2 (effectively replicating results of Experiment 1; see Figure 1). In short, it seems that in putting forth a valiant attempt to avoid discriminatory bias, participants drastically reduced the severity of their sentences; however, they did so for both Latino and White defendants, leaving the differential bias intact.

General Discussion

The present results have numerous implications for current models of prejudice and attitude-behavior correspondence. Foremost, this study reveals the paradoxical finding that even well-considered, deliberative, ostensibly controllable responses can be inconsistent with one's intent or motivation to respond in a particular manner. While prior research has defined deliberative behavior in terms of cognitive resources, control, or time constraints (Fazio, 1990), the present data suggest that current conceptualizations of "deliberative" or "control" be modified to include epistemic as well as agentic factors. As Dovidio and colleagues point out, "the spontaneous-deliberative disctinction requires further conceptual refinement that identifies the factors…that critically define behaviors as deliberative" (Dovidio et al., 1997; p. 536). For instance, individuals may literally be <u>unable</u> to response in a nonprejudiced manner on ostensibly deliberative tasks when they are unsure what a nonprejudiced response entails, even if they have sufficient time and resources to consider various options.

Perhaps a subtle but important distinction should be drawn between control over behavior per se and control over discriminatory outcomes. The latter may depend as much on epistemic

limitations as mere agency per se. An example given by Devine and Vasquez (1999) entails a highly motivated, nonprejudiced individual who is having a pizza party and is unsure about inviting a certain foreign guest because she does not want to offend her by consuming food that might violate this guest's religious beliefs. Here the behavior is highly controlled (i.e., the act of inviting a guest), but cultural ignorance, in this case, creates a situation in which an well-intentioned, controllable action could yield a discriminatory outcome (e.g., failing to invite a guest, based on her religion, to a party that she would have otherwise enjoyed). Similarly, participants in the present study are put in a situation where they have to punish a criminal convicted of a heinous crime, without being racially biased. Although the sentencing process is highly controlled and deliberate, participants must possess some knowledge of what a biased response entails in order to avoid a discriminatory outcome. Although there are clear wrong answers—verbal reprimand would be too lenient, and that death would be too severe (no participant selected either response), there is no clear right answer. Should the defendant receive probation or prison time? How much? This response ambiguity creates a situation that is ripe for the unwanted influence of racial attitudes on decision outcomes. However, in most tests of the MODE model, the experimental paradigm is set up to contain a "correct" response, which participants can surmise if they have the motivation and are given the opportunity to do so (e.g., whether to buy a camera from Smith's or Brown's department store; see Sanbonmatsu and Fazio, 1990).

In a similar vein, motivation itself may be a multidimensional construct. For instance, there may be a qualitative distinction between the motivation to be <u>fair</u> and the motivation to be <u>accurate</u>. While the motivation to be accurate inherently assumes that there is a correct response, the motivation to be fair only requires careful consideration and unbiased intent. The present

study emphasized the motivation to be fair whereas the MODE model has focused on the motivation to be accurate . As Fazio and Towles-Schwen (1999) state, “the MODE model tends to focus on a broad motivation to be accurate…however, we do recognize that the motivation to deliberate can also stem from more specific goals regarding the standards that individuals maintain for their behavior in a given domain” (p. 100).

Two other issues that arise concern questions of when unintentional discrimination occurs and whether such bias can be overcome. The present results and past research indicate that several situational factors may affect the occurrence of unintentional discrimination. As previously discussed, one such factor is response ambiguity. That is, unintentional discrimination is more likely to occur when response options or norms regarding socially appropriate behavior are vague (Dovidio & Gaertner, 1996; Gaertner & Dovidio, 1986; Sargent & Theil, 2002). In such situations, motivation is less effective in overriding biased processing generated by racial attitudes. A second factor is stereotypic fit. Discriminatory bias is more likely to emerge when there is a correspondence between the stereotype associated with the target and the action performed by the target (Bodenhausen, 1988; Bodenhausen & Lichtenstein, 1987). In a similar vein, research in the psycholegal literature has shown that racial bias depends on the nature of the crime (Gordon et al., 1988; Sunnafrank & Fontes, 1983). Minority defendants who perpetrate counterstereotypic, white-collar crimes, such as securities fraud or embezzlement may actually receive less severe sentences than Whites (Mazella & Feingold, 1994). Consistent with this notion, when participants were randomly assigned to read about random violent assaults against a stranger (i.e., the crime scenarios in the present studies) or assaults in the context of domestic violence, which is not more stereotypically associated with minorities (Esqueda, 1997), discriminatory bias was obtained for the random assault but not domestic assault (Livingston,

2001). In brief, unintentional discriminatory bias may only occur in certain contexts, namely when there is response ambiguity and when there is strong stereotypic fit between target ethnicity and target behavior.

The mental contamination model outlines a number of conditions that must be satisfied in order to correct for unwanted bias when it does emerge, among which is the ability to estimate the magnitude of bias. Even if individuals were motivated to avoid bias, the response ambiguity involved with the sentencing task may have obfuscated the magnitude of the bias. During debriefing discussions, most participants seemed sincere in their disavowal of racial discrimination, insisting that the severity of their sentences was based entirely on the severity of the crime itself, and would have been no different for a White defendant. In addition, individuals must also be psychologically able to adjust for bias. Revisiting the teacher illustration, "Jones may know that biased processing has led to a lowering of Hernandez's grade from a B to a C but may simply be unable to escape the impression that the paper is truly mediocre" (Wilson & Brekke, 1994). Similarly, much of the discrimination may reflect a cognitive bias in which attitudes distort participants' view the crime itself and the defendant. In such cases, White participants may simply be unable to avoid perceiving a violent crime committed by a minority defendant as being more morally reprehensible than the identical crime committed by a White defendant.

The results of the present studies reveal grim prospects for correcting unintentional bias by using motivational tactics. Results from Experiment 1 indicate that notifying individuals of the potential for bias and encouraging them to be fair had absolutely no effect on discriminatory bias. Even the emphatic, overblown attempts to raise motivation and eliminate bias in Study 3 were unsuccessful. Although these findings might seem surprising or unexpected on the surface,

they make sense if (lack of) motivation is not the source of discriminatory bias. In other words, if the underlying antecedent of unintentional bias is cognitive rather than motivational, it is only natural that motivation-enhancing manipulations would be ineffective in reducing bias. As previously mentioned, inability to psychological adjust one's perceptions of an event may underlie discriminatory bias. Perhaps strategies designed to reduce negative stereotypic associations between minorities and violent crime or to alter negative evaluative associations would be more successful in eliminating bias (Blair & Banaji, 1996; Dasgupta & Greenwald, 2001). In sum, the present study shows that the road to deliberative discrimination can be paved with good intentions. Deliberative discrimination does not necessarily depend on motivational processes, but rather can occur unintentionally as the result of mental contamination (Wilson & Brekke, 1994). The present findings indicate that malicious intent is hardly a sine qua non for the occurrence of collective discrimination, and may reconcile the seemingly contradictory notions that (1) contemporary society is egalitarian, while (2) systemic discrimination against stigmatized minority groups continues to persist.

References

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. Personality and Social Psychology Review, 6, 242-261.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. Journal of Personality and Social Psychology, 70, 1142-1163.

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. Journal of Personality and Social Psychology, 55, 726-737.

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information processing strategies: The impact of task complexity. Journal of Personality and Social Psychology, 52, 871-888.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. Journal of Personality and Social Psychology, 82, 359-378.

Dasgupta, N., & Greenwald, A. G. (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.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology, 56, 5-18.

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping. In S. Chaiken & Y. Trope (Eds.), Dual process theories in social psychology (pp. 339-360). New York: Guilford.

Devine, P. G., & Vasquez, K. A. (1998). The rocky road to positive intergroup relations. In J. L. Eberhardt & S. T. Fiske (Eds.), Confronting racism: The problem and the response (pp. 234-262). Newbury Park, CA: Sage.

Dovidio, J. F., & Gaertner, S. L. (1996). Affirmative action, unintentional racial biases, and intergroup relations. Journal of Social Issues, 52, 51-75.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. Journal of Personality and Social Psychology, 82, 62-68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). The nature of prejudice: Automatic and controlled processes. Journal of Experimental Social Psychology, 33, 510-540.

Esqueda, C. W. (1997). European American students' perceptions of crimes committed by five racial groups. Journal of Applied Social Psychology, 27, 1406-1420.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 23, pp. 75-109). San Diego, CA: Academic Press.

Fazio, R. H., Jackson, J., Dunton, & Williams (1995). Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? Journal of Personality and Social Psychology, 69, 1013-1027.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. Journal of Personality and Social Psychology, 50, 229-238.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp. 97-116). New York: Guilford Press.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), Prejudice, discrimination, and racism (pp.61-89). Orlando, FL: Academic Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74, 1464-1480.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system justification and the production of false consciousness. British Journal of Social Psychology, 33, 1-27.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), Attitude strength: Antecedents and consequences (pp. 1-24). Mahwah, NJ: Erlbaum.

Kruglanski, A. W. & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. Journal of Experimental Social Psychology, 19, 448-468.

Landy, D., & Aronson, E. (1969). The influence of the character of the criminal and his victim on the decisions of simulated jurors. Journal of Experimental Social Psychology, 5, 141-152.

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

Livingston, R. W. (2001). Bias in the absence of malice: The phenomenon of unintentional discrimination. Unpublished doctoral dissertation, The Ohio State University.

Mazella, R. & Feingold, A. (1994). The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta-analysis. Journal of Applied Social Psychology, 24, 1315-1344.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), Prejudice, discrimination, and racism (pp. 91-125). Orlando, FL: Academic Press.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. Journal of Personality and Social Psychology, 75, 811-829.

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. Journal of Personality and Social Psychology, 59, 614-622.

Sargent, M. J., & Theil, A. (2002). Assessing the predictive utility of the "Black-White" Implicit Association Test: Evidence for moderators. Paper presented at the annual meeting of the Society for Personality and Social Psychology, Savannah, GA.

Sidanius, J., & Pratto, F. (1999). Social dominance. Cambridge University Press.

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

Swanson, J. E., Rudman, L. A., & Greenwald, A. G. (2001). Using the Implicit Association Test to investigate attitude-behavior consistency for stigmatized behaviour. Cognition and Emotion, 15, 207-230.

Sweeney, L. T., & Haney, C. (1992). The influence of race on sentencing: A meta-analytic review of experimental studies. Behavioral Sciences and the Law, 10, 179-195.

Vargas, P. T., von Hippel, W., Petty, R. E. (2002). Types of information processing and attitude behavior relations. Paper presented at the annual meeting of the Society for Personality and Social Psychology, Savannah, GA.

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

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

Table 1

Correlations between discrimination and implicit prejudice, explicit prejudice, and motivation to control prejudice as a function of defendant ethnicity (Experiment 1)

Measure	**White**	**Latino**
Implicit prejudice	-.19	.37+
Explicit prejudice	.04	.10
Motivation to avoid prejudice	.17	-.05

Note.
+ $p < .10$

Table 2

Correlations between discrimination and implicit prejudice, explicit prejudice, and motivation to control prejudice as a function of defendant ethnicity (Experiment 2)

Measure	**White**	**Hispanic**	**Black**
Implicit prejudice	-.03	.04	.43*
Explicit prejudice	.07	-.29	.70**
Motivation to avoid prejudice	.33^	-.24	-.32+

Note.
*p < .05
**p < .01
+ p = .08
^p = .12

Table 3

Regression for attitudes or motivation as predictors of discrimination against Latino or Black defendant vis-à-vis White defendant (Experiment 2)

	Latino		Black		Latino		Black		Latino		Black	
Variables	b	p	B	P	B	P	B	P	B	p	B	P
Race	**.54**	.007	**.38**	.03	.26	.21	-.01	.96	**.29**	.03	**.28**	.03
Explicit Prejudice	.04	.83	.15	.33								
Race x Explicit	-.27	.20	**.44**	.01								
Implicit Prejudice					.02	.92	-.02	.92				
Race x Implicit					.06	.24	.48	.08				
Motivation									.26	.22	.29	.25
Race x Motivation									-.40	.06	**-.51**	.04

Table 4

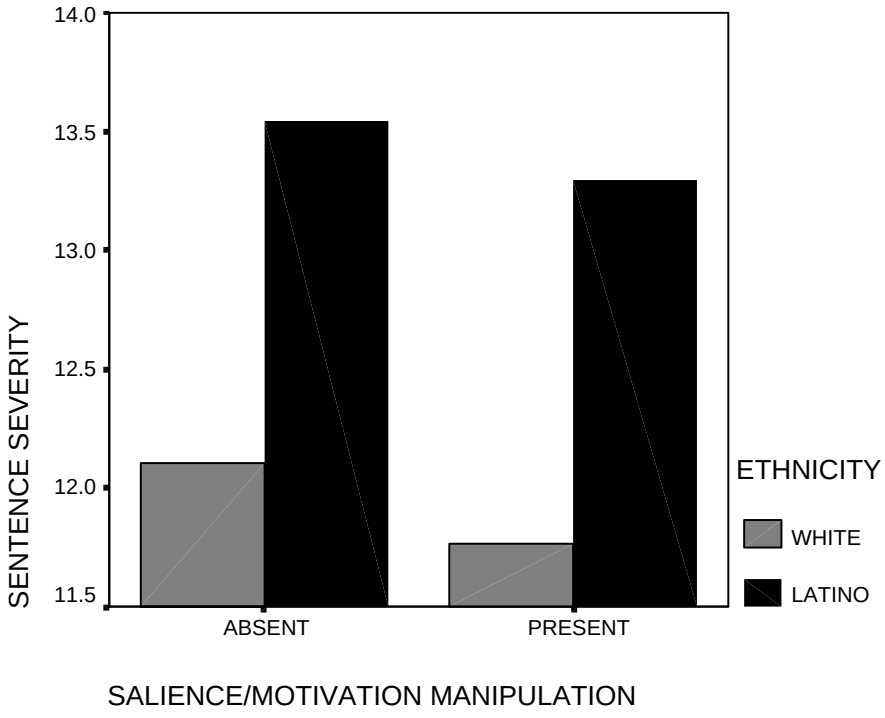
Simultaneous regression for variables predicting deliberative discrimination against Black defendant (Experiment 2)

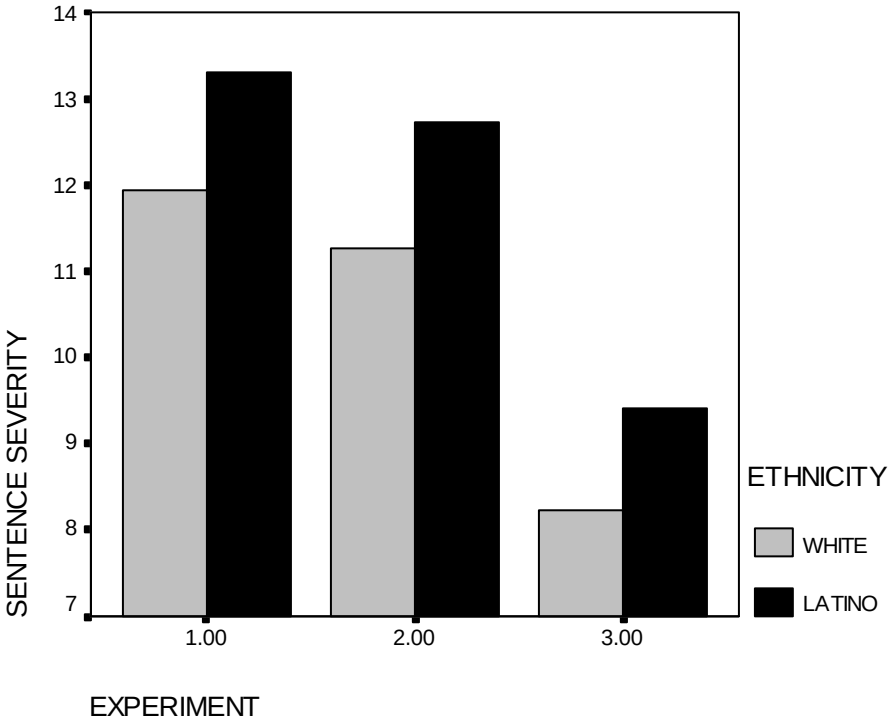
Equation	1		2	
Variables	b	P	b	p
Race	.02	.94	.25	.19
Explicit Prejudice			.10	.51
Race x Explicit Prejudice			**.60**	.006
Implicit Prejudice	-.04	.84		
Race x Implicit Prejudice	.44	.14		
Motivation to Control Prejudice	.29	.24	.24	.29
Race x Motivation	-.34	.20	.05	.20

Figure Captions

Figure 1. Sentence severity as a function of defendant ethnicity and experimental condition. Experiment 1.

Figure 2. Sentence severity as a function of defendant ethnicity. Experiments 1, 2, and 3.


14.0
13.5
13.0
12.5
12.0
11.5
SENTENCE SEVERITY
ABSENT
PRESENT
SALIENCE/MOTIVATION MANIPULATION
ETHNICITY
WHITE
LATINO


14
13
12
11
10
9
8
7
SENTENCE SEVERITY
1.00
2.00
3.00
EXPERIMENT
ETHNICITY
WHITE
LATINO

Author Note

This manuscript is based, in part, on a dissertation submitted to The Ohio State University in partial completion of the requirements for the doctorate degree. This dissertation was the recipient of a 2002 SPSSI Social Issues Dissertation Prize. The author wishes to thank members of the dissertation committee: Rich Petty, Bill von Hippel, Lisa Flores, and particularly Marilynn Brewer. Correspondence concerning this manuscript should be addressed to Robert Livingston, Department of Psychology, 1202 W. Johnson Street, Madison, WI 53706, or via e-mail at rwlivingston@facstaff.wisc.edu.

Footnotes

[1] The mental contamination model does not specify that the attitudes themselves be nonconscious, only that the effect of the attitude on subsequent behavior be nonconscious.

[2] Because the target was Hispanic, items were modified to assess motivation to be nonprejudiced against minorities in general (e.g., I am personally motivated by my beliefs to be nonprejudiced toward minorities), as opposed to Blacks specifically.

[3] The Sommers and Ellsworth (2000) study which found an effect for race salience used Black participants. As will be tested in the next study, the race of the participant may have an effect on the nature of findings. Additionally, Sommers and Ellsworth found more consistent and stronger findings for guilt attribution rather than criminal sentencing across their studies. Prior research has shown that there may exist qualitative differences between the processes involved with guilt attribution and criminal sentencing (Sweeney & Haney, 1992). Finally, consistent with other studies (e.g., Livingston, 2001; Mazella & Feingold, 1994), the data from Sommers and Ellsworth indicated that the nature of crime itself influenced the degree of racial bias (see Sommers & Ellsworth, 2000; Study 1). In short, there are many situational and methodological factors that make it difficult to compare findings from different studies that use different mock juror paradigms, crimes, or target characteristics.

[4] Due to a procedural oversight, feeling thermometer data were collected for only 48 participants. However, this error in data collection was not differential across the three ethnicity conditions.

[5] One interesting observation is that motivation to control prejudice was asymmetrically related to bias against minorities and whites. That is, high motivation to control prejudice was associated with less severity against minorities while it was associated with more severity toward White defendants. Crandall et al. (2002) have argued that such patterns indicate that motivation scales do not tap a general motivation to be nonprejudiced per se but rather specific awareness of and sensitivity to social norms that proscribe prejudice against certain groups (but may actually encourage prejudice toward other groups). Consistent with this idea, motivation to control prejudice was significantly related to the suppression of discrimination against Blacks but only marginally related to the suppression of prejudice against Latinos, a group for whom nonprejudiced norms are weaker (Crandall et al., 2002).

[6] The somewhat weaker relationship between discrimination and the IAT for Hispanics in Study 2 compared with Study 1 could have been due to the fact that implicit measures were administered during the same session, and subsequent to the explicit measures and the sentencing task. This may have heightened reactivity to the IAT. Past research has shown that motivation can affect implicit measures (see Blair, 2002 for review). Consistent with this idea, motivation to control prejudice and implicit prejudice were not correlated for participants in Study 1 where the IAT was administered in a separate session $r = -.03$, $p < .83$, however they were correlated for participants in Study 2, $r = -.35$, $p < .001$ when the IAT was given in the same session. This suggests that participants may have suppressed implicit prejudice in Study 2, implying that the relationship between the IAT and discrimination against Blacks found in Study 2 might have been even stronger if the IAT were measured in a separate experimental session.

[7] It was decided a priori that these manipulations would be administered to all participants rather than doubling the needed sample size by including a control condition. This decision seemed reasonable at the time given that results from the first two studies indicated that bias surely would emerge in a control condition, and the main goal of Experiment 3 was to test whether bias would not emerge.