Research Article

Psychological Science
201X, Vol XX(X) 1–9

DOI: 10.1177/0956797613501171
pss.sagepub.com

Stereotyping to Infer Group Membership Creates Plausible Deniability for Prejudice-Based Aggression

William T. L. Cox and Patricia G. Devine
University of Wisconsin–Madison

Abstract

In the present study, participants administered painful electric shocks to an unseen male opponent who was either explicitly labeled as gay or stereotypically implied to be gay. Identifying the opponent with a gay-stereotypic attribute produced a situation in which the target's group status was privately inferred but plausibly deniable to others. To test the *plausible deniability hypothesis*, we examined aggression levels as a function of internal (personal) and external (social) motivation to respond without prejudice. Whether plausible deniability was present or absent, participants high in internal motivation aggressed at low levels, and participants low in both internal and external motivation aggressed at high levels. The behavior of participants low in internal and high in external motivation, however, depended on experimental condition. They aggressed at low levels when observers could plausibly attribute their behavior to prejudice and aggressed at high levels when the situation granted plausible deniability. This work has implications for both obstacles to and potential avenues for prejudice-reduction efforts.

Keywords

prejudice, aggression, violence, stereotypes, stereotyping, stereotyped attitudes

Received 4/19/13; Revision accepted 7/21/13

Although many people consider prejudice to be immoral and socially unacceptable, it persists throughout society. Rates of prejudice-based harassment, discrimination, bullying, and hate crimes worldwide are alarmingly high (Bleich, 2011; Katz-Wise & Hyde, 2012; U.S. Department of Justice, 2010). The U.S. media, tracking many recent instances of extreme prejudice-based bullying, has declared a bullying epidemic (e.g., Cooper, 2011; Couric, 2012). As noted by Cox, Abramson, Devine, and Hollon (2012), this "epidemic" is especially disconcerting because prejudice-based violence can cause severe mental-health issues, drive the victims to commit suicide (e.g., Tyler Clementi, Carl Walker-Hoover), or even escalate to homicide (e.g., James Byrd, Lawrence King, Matthew Shepard). The impact of prejudice-based violence extends beyond the victims and their families, terrorizing and intimidating whole communities. Understanding why and how such blatant prejudice continues and how to reduce it has been a central concern of social psychologists for decades.

Powerful social norms and institutional policies have been developed to discourage the expression of prejudice (e.g., hate-crime laws, antibullying policies). When prejudice is expressed toward someone whose group status is publicly known, these external pressures can easily be applied. Violence against a Black man, for example, is easily attributable to prejudice based on the victim's race, and observers can therefore enforce anti-prejudice norms against the perpetrator. Many stigmatized groups, however, lack visible defining features (e.g., gay men, Jewish people). Nevertheless, people often use stereotypes to make assumptions about others' membership in these groups (e.g., inferring that a fashionable man is gay; Matthews & Hill, 2011; Shelp, 2002). These assumptions about group membership, whether correct

Corresponding Author:
William T. L. Cox, Department of Psychology, University of Wisconsin–Madison, 1202 W. Johnson St., Madison, WI 53706
E-mail: will.cox@me.com

or incorrect, may lead a perpetrator to express prejudice. Because observers lack access to the perpetrator's private assumption about group membership, they cannot confidently attribute his or her negative behavior to prejudice. Such situations, therefore, grant perpetrators "plausible deniability" for their prejudicial motives, which frees them from external pressures that would normally discourage prejudice. In the present work, we tested this *plausible deniability hypothesis* in the context of prejudice-based aggression.

Plausible deniability involves strategically evading external pressures from others, in contrast to hiding prejudicial behavior from the self (as in aversive prejudice[1]; Gaertner & Dovidio, 1986). Testing the plausible deniability hypothesis, therefore, requires identifying people whose personal convictions do not oppose prejudice but who are sensitive to social pressures that oppose prejudice. Individual differences in sensitivity to personal versus social standards for responding without prejudice have been explored in depth by Plant and Devine (1998, 2001, 2009) and are key to predicting who will express prejudice and when they will do so. Internal motivation to respond without prejudice reflects a person's personal values and standards, whereas external motivation to respond without prejudice reflects a person's sensitivity to societal pressures and sanctions that oppose prejudice. People can be high or low in either or both types of motivation, and the combination of motivations determines whether and under what circumstances people will express prejudice.

A situation's plausible deniability should be immaterial to people who are high in internal motivation to respond without prejudice, because being nonprejudiced is personally important to them. People with low levels of internal motivation to respond without prejudice, however, are generally more likely to express prejudice because doing so does not violate their personal values (Plant & Devine, 1998, 2001, 2009). Among people who lack internal motivation to respond without prejudice, whether and when they express prejudice depends on their level of external motivation to respond without prejudice and the situational relevance of social pressures (e.g., whether plausible deniability is present or absent). If people are low in both internal and external motivation to respond without prejudice, they are likely to express prejudice whether plausible deniability exists or not, because they disregard what others think. In contrast, if people are low in internal but high in external motivation, and therefore want to appear nonprejudiced to others, they will refrain from expressing prejudice when external standards are relevant. These *low-internal, high-external* people strategically hide prejudice from others, but freely express prejudice when there are no chances of "being caught." It is these low-internal, high-external people, therefore, who should express prejudice in situations that grant plausible deniability, because they can evade external pressures and negative judgment from others.

The Present Experiment

We tested the plausible deniability hypothesis in a laboratory aggression experiment. We measured people's internal and external motivations to respond without prejudice and provided an opportunity for aggression. Participants believed they were giving painful electric shocks to another person. In the plausible deniability absent condition, the target's stigmatized group status was publicly known, thus making external standards relevant to behavior toward the target. In the plausible deniability present condition, however, the target's stigmatized group status was implied by a stereotypic trait, thus granting plausible deniability. Two control conditions addressed possible alternate explanations.

The plausible deniability hypothesis involves a three-way interaction among plausible deniability (absent vs. present), internal motivation, and external motivation. We expected that across plausible deniability conditions, people high in internal motivation would aggress at low levels, and people low in both internal and external motivation would aggress at high levels. We expected that aggression levels would vary by plausible deniability condition only for the people who are low in internal motivation and high in external motivation. When plausible deniability was absent, these people would hide their prejudice and aggress at low levels, but when plausible deniability was present, they would aggress at high levels because they could express their personally held prejudice without fear of social sanction.

Method

Participants and design

Heterosexual undergraduates (N = 166; 102 men, 64 women[2]) participated in exchange for course credit. Participants' levels of internal and external motivation to respond without prejudice were assessed several weeks before the experimental session. In the experimental session, participants were randomly assigned to one of four conditions (plausible deniability present or absent, or one of two control conditions) and were given the opportunity to aggress toward an opponent in an aggression paradigm.

Internal and external motivations to respond without prejudice

In an online mass prescreening survey, among a large set of other questionnaires, participants completed measures

of internal motivation to respond without prejudice (IMS) and external motivation to respond without prejudice (EMS), modified for prejudice against gay men (Plant & Devine, 1998). The IMS items measure motivation to respond without prejudice for internal, personal reasons (e.g., "I am personally motivated by my beliefs to be nonprejudiced toward gay men"), whereas the EMS items measure motivation to respond without prejudice for external, social reasons (e.g., "I try to hide any negative thoughts about gay men in order to avoid negative reactions from others"). There are 5 IMS items ($\alpha = .9$) and 5 EMS items ($\alpha = .8$), each scored on a Likert scale from 1 (*strongly disagree*) to 9 (*strongly agree*). Participants were sampled to ensure that we obtained the full conceptual range of IMS and EMS (IMS range = 1–9; EMS range = 1–8.2).

Selection of the stereotypic attribute

Testing our central hypothesis required that IMS and EMS be related to the expression of prejudice but unrelated to relying on a stereotype to infer group membership. Therefore, we needed to be highly confident a priori that participants, regardless of their levels of IMS and EMS, would reliably infer that a man possessing the stereotypic attribute was gay.

In a separate online study, 50 undergraduate students, none of whom were in the main experiment, were asked to list the first thoughts that "popped to mind" about "a man who likes shopping." Thirty-six (72%) stated that the man was gay (e.g., they wrote "gay" or "homosexual"). Five of the 14 remaining participants supplied responses that we coded as implying that the man was gay (e.g., "feminine," flamboyant"). Therefore 82% of participants stated or implied that the man was gay. As in the main experiment, these participants represented a full range of IMS scores (1.40–9, $M = 6.65$, $SD = 1.973$; $\alpha = .9$) and EMS scores (1–9, $M = 4.73$, $SD = 1.885$; $\alpha = .8$). IMS, EMS, and their interaction were unrelated to whether participants stated that the man was gay, and were unrelated to whether participants stated or implied that the man was gay ($|\text{pairwise } rs| \leq .03$, $ps \geq .832$; $|\beta s| \leq 0.05$, $ps \geq .770$). We concluded, therefore, that this phrase would reliably lead participants in the experiment to infer that a man who likes shopping is gay, regardless of their levels of IMS and EMS.

Manipulation and procedure

Participants were led to believe that they were competing with an unseen male opponent. They received two extra-credit points in their introductory psychology class in exchange for participating and believed they were competing with the opponent to win a third point (however, all participants received three points after debriefing). Participants learned that the opponent was male, and they received one "identity statement" that he purportedly provided, which created the opportunity to introduce the manipulation. In the plausible deniability absent condition, the information explicitly stated that the opponent was gay ("I am gay"; $n = 40$). In the plausible deniability present condition, it expressed a gay-stereotypic attribute that reliably leads people to infer that a man is gay ("I like shopping"; $n = 42$).

As noted previously, the plausible deniability hypothesis states that aggression will vary as a function of the presence or absence of plausible deniability and internal and external motivations to respond without prejudice. As measured variables, however, the motivations to respond without prejudice may correlate with third variables (e.g., general aggressiveness), which would undermine the conclusions that can be drawn. To address such alternatives, we included two control conditions in which participants believed that the opponent was straight. In the first control condition, the opponent declared he was straight ("I am straight"; $n = 41$). This straight control condition allowed us to test whether internal and external motivations were related to aggression generally rather than to prejudice-based aggression specifically. In the second control condition, the opponent liked shopping but was known to be straight ("I like shopping with my girlfriend"; $n = 43$). This shopping control condition allowed us to test whether something specific about liking shopping (e.g., gender deviance, consumerism) would create a relationship between aggression and the motivations to respond without prejudice that was independent of an inference that the opponent was gay. If the aggression observed in the plausible deniability conditions reflected antigay prejudice, rather than a third variable, internal and external motivations to respond without prejudice would be unrelated to aggression in either control condition.

We selected experimenters that we judged would be seen as White and heterosexual. Experimenters wore professional, business-casual clothing and white lab coats. Two experimenters conducted each experimental session, and the experimenters' gender always matched the participant's gender.

Experimenter 1 explained the procedure and told the participant that we were studying how "isolated pieces of personal information" influence impression formation. To that end, participants generated 20 self-descriptive identity statements. We required a large number of statements to make it plausible that mundane information would be among the opponent's statements (e.g., "I like shopping").

As Experimenter 1 connected the shock apparatus to the participant's ankle, Experimenter 2 arrived and announced that the opponent was ready. During this

interaction, Experimenter 2 mentioned twice that the opponent was male, gave Experimenter 1 a bowl containing the opponent's 20 identity statements, and then left the room. The bowl contained a mix of the 4 statements that represented the four conditions. The participant was instructed to select one statement from the bowl, without looking, and to then pin it on a corkboard in front of him or her. The rigged identity statements were all written by the same research assistant, in neat, legible handwriting that we judged to be not especially gay, straight, masculine, or feminine.

While completing a filler questionnaire, participants heard Experimenter 1 using a paper cutter to separate their statements in the adjacent desk area and then saw Experimenter 1 put the statements into a bowl identical to the one that contained their opponent's statements. Participants then saw Experimenter 1 leave the room with the bowl and heard him or her knock on a door across the hall. Experimenter 1 gave the bowl to Experimenter 2, and then the two experimenters had a short discussion, in which they arranged to connect their computer systems for the interaction. Participants could readily hear this conversation.

Participants then completed the shock task. Throughout the shock task, Experimenter 1 was in contact with the participant. Experimenter 1 indicated that he or she was monitoring the session and was ready to respond to either questions or any signs of distress as the shocks were being delivered. Therefore, the participant was aware that Experimenter 1 was monitoring the exchange of shocks, which made him or her a salient audience to the participant's behavior. After the shock task, participants completed a debriefing questionnaire, were awarded extra credit, and left the room.

Shock task

Because people's sensitivity to electric shocks is highly variable, the appropriate intensity levels were established separately for each participant. Participants received shocks of increasing intensity and were asked to rate each shock on a scale from 0 to 20 (0 = *can't feel it*, 1 = *tangible*, 5 = *uncomfortable*, 10 = *unpleasant*, 15 = *disagreeable*, 20 = *most you can tolerate*). The goal was to identify the level of shock that was "uncomfortable" and the level that was "the most they could tolerate," which were subsequently used as the endpoints (converted to 1 and 9) on a continuum of shock intensity received during the task. This also established that the shock apparatus was real and enhanced the cover story.

The task was adapted from the Taylor (1967) aggression paradigm (see also Berkowitz, 1989; Donnerstein, Donnerstein, Simon, & Ditrichs, 1972; Giancola & Parrott, 2008; Zeichner, Frey, Parrott, & Butryn, 1999). On each trial, participants watched a yellow circle on their computer monitor until, after a random interval, it turned red. When the circle became red, the participants' goal was to turn the knob on their response apparatus more quickly than their opponent did. They completed two practice trials (one win and one loss) to become familiar with the task.

After the practice trials, the task had 21 trials, rigged so that participants won 11 trials. On each winning trial, they shocked their opponent. They selected the intensity of the shocks on a 9-point Likert scale (1 = *uncomfortable*, 9 = *the most pain they can tolerate*) and the duration of the shocks (50, 100, 200, 250, or 500 ms). On losing trials, participants received shocks of the intensity and duration they had selected previously (e.g., on their first loss, they received the shock they sent out from their first win, and so on). To accomplish this, we constructed a pattern of wins and losses that always gave the participants some wins before losses (e.g., win, loss, win, win, loss, win, loss, loss). All participants experienced the same win/loss pattern. Rigging the win/loss pattern in this way standardized the reciprocation of aggression (e.g., if participants gave a low shock, the "opponent" never responded with a disproportionately high shock, as could happen if the shocks' intensity and duration were random). Participants were not told the exact intensity and duration of the shocks they received.[3]

Aggression metric

Because shock intensity and duration constitute two separate ways participants might aggress on each trial, we combined the two measures into a single metric. We standardized the duration and intensity responses, recoding them as percentages of the greatest possible shock (500 ms or 9, "the most you can possibly handle," was coded as 100%; 50 ms or 1, "uncomfortable," was coded as 0%). Our metric, the average of these intensity and duration responses across all 11 winning trials ($\alpha = .95$), reflected the percentage of the maximum amount of shock it was possible to give the opponent. Selecting the highest intensity and longest duration for every shock (the maximum amount of aggression possible) would yield a score of 100%. Selecting the lowest intensity and shortest duration for every shock (the minimum amount of aggression possible) would produce a score of 0%. When analyzed separately, duration and intensity yield patterns and effect magnitudes that match those reported in the Results section using the composite metric.

Results

IMS, EMS, and aggression scores are reported in Table 1. IMS and EMS were mean-centered in all regression analyses.

Table 1. Means and Standard Deviations in Each Condition

	Plausible deniability conditions		Control conditions	
Measure	Absent (opponent was gay)	Present (opponent liked shopping)	Straight (opponent was straight)	Shopping (opponent liked shopping with his girlfriend)
IMS	6.24 (1.873)	6.37 (2.441)	6.77 (2.009)	6.38 (1.856)
EMS	4.41 (1.764)	4.16 (1.837)	4.04 (1.660)	4.33 (1.774)
Aggression	20.3 (20.83)	18.0 (24.41)	18.5 (21.65)	14.4 (18.89)

Note: Standard deviations are given in parentheses. IMS and EMS reflect levels of internal and external motivation to respond without prejudice, respectively, described in Plant and Devine (1998). Aggression reflects the percentage of the maximum amount of shock it was possible to administer to the opponent.

Plausible deniability hypothesis

We conducted a regression analysis to determine whether plausible deniability (absent = 0, present = 1), IMS, EMS, and their two- and three-way interactions predicted aggression.[4] The main effect of IMS ($\beta = -0.387$, $p = .023$) and the IMS × EMS interaction ($\beta = 0.374$, $p = .020$) were qualified by the predicted Plausible Deniability × IMS × EMS interaction ($\beta = -0.420$, $p = .011$). There were no other main effects or interactions, $|\beta s| \leq 0.217$, $ps \geq .144$. We explored the three-way interaction by conducting separate regression analyses within each plausible deniability condition.

In the plausible deniability absent ("gay") condition, when others could readily attribute negative behavior to prejudice, there was no main effect of EMS ($\beta = -0.139$, $p = .322$), and the main effect of IMS ($\beta = -0.362$, $p = .023$) was qualified by an IMS × EMS interaction ($\beta = 0.373$, $p = .011$). This interaction revealed that only participants who were low in both IMS and EMS aggressed at high levels (Fig. 1a). This pattern matches the findings of previous work (e.g., Plant & Devine, 2001) that has explored IMS and EMS and a target whose group status was publicly known and unambiguous (e.g., race). Although their rates of aggression are

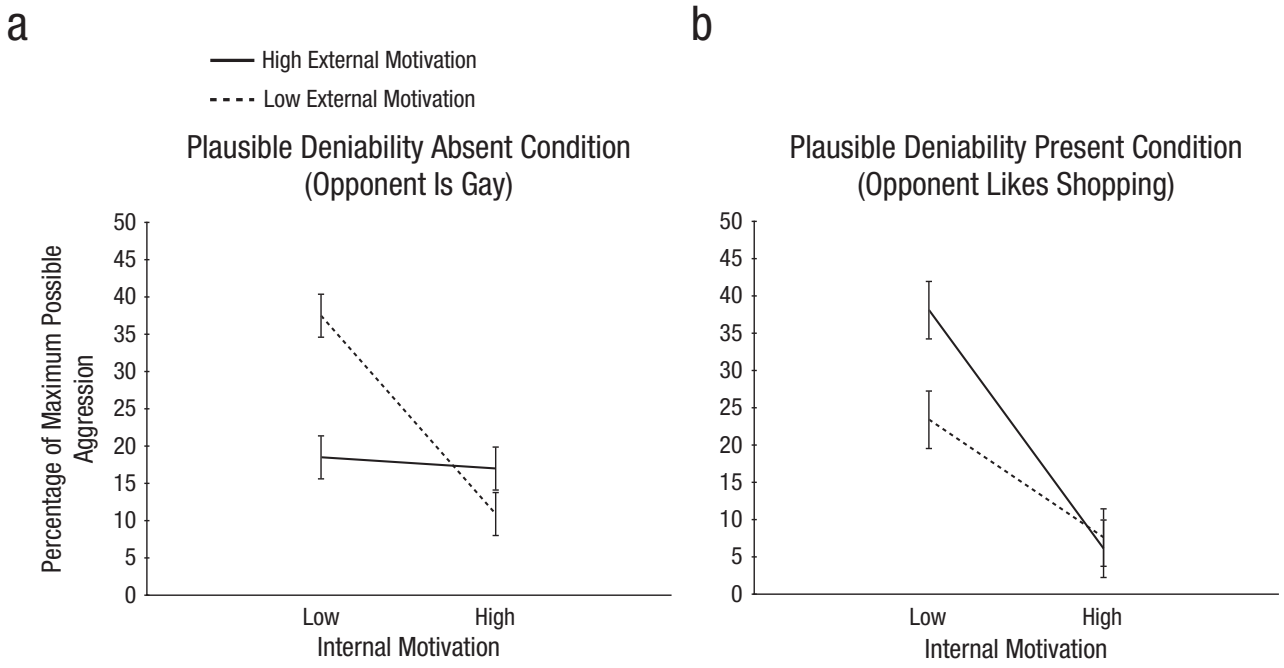

Fig. 1. Results for the plausible deniability conditions. The percentage of the maximum possible aggression, a composite of intensity and duration of electric shock administered, is shown as a function of internal and external motivation to respond without prejudice. Results are shown separately for conditions in which plausible deniability was absent (a) and in which it was present (b). High and low internal and external motivation were plotted at 1 standard deviation above and below their respective means. Error bars represent the standard errors of the mean level of aggression within condition.

similar, the underlying motivations of high-IMS people and low-IMS, high-EMS people are different. Whereas high-IMS people personally renounce prejudice, low-IMS, high-EMS people strategically refrain from aggressing because others could attribute such behavior to prejudice.

In the plausible deniability present ("shopping") condition, in which orientation was implied by a stereotype, a different pattern emerged. Specifically, only IMS predicted the level of aggression ($\beta = -0.570$, $p = .001$): Low-IMS participants aggressed at much higher levels than high-IMS participants. Neither EMS nor IMS × EMS ($|\beta s| \leq 0.173$, $ps \geq .259$) predicted aggression (Fig. 1b). Concerns about social opposition to prejudice (i.e., external motivation) were irrelevant in this condition, and only personal values (i.e., internal motivation) related to aggression.

In short, only low-IMS, high-EMS people behaved differently across the plausible deniability conditions. They aggressed at low levels when their behavior could be plausibly attributed to prejudice and at high levels when the situation granted plausible deniability. Stereotyping to infer group membership grants plausible deniability, and plausible deniability undermines external, but not internal, motivation to respond without prejudice.

Control conditions

The straight control condition allowed us to assess whether IMS and EMS, as measured individual differences variables, correlated with aggression generally rather than prejudice-based aggression specifically. Eliminating this alternate explanation, we found that rates of aggression in the straight control condition were unrelated to IMS, EMS, or the IMS × EMS interaction ($|\beta s| \leq 0.182$, $ps \geq .302$), as shown in Figure 2a.

The shopping control condition allowed us to assess whether the opponent's liking shopping, rather than an inference about his being gay, created a relationship between aggression and IMS in the plausible deniability present condition. Opposing this possibility, we found that rates of aggression in the shopping control condition were unrelated to IMS, EMS, or their interaction ($|\beta s| \leq 0.227$, $ps \geq .149$), as shown in Figure 2b.

The results from these control conditions assure us that the patterns of aggression related to IMS and EMS in the plausible deniability conditions reflected antigay aggression specifically, rather than being an artifact of third variables related to our individual differences measures or the specific attribute we used to imply the opponent's group membership.

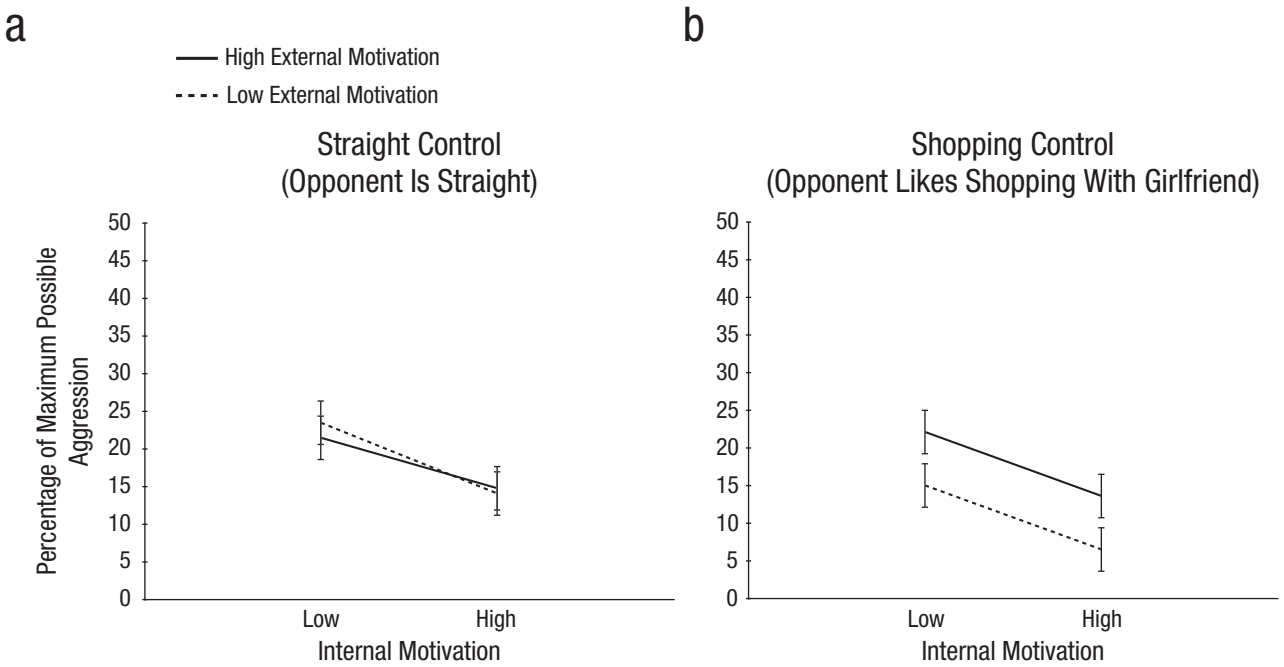

Fig. 2. Results for the control conditions. The percentage of the maximum possible aggression, a composite of intensity and duration of electric shock administered, is shown as a function of internal and external motivation to respond without prejudice. Results are shown separately for conditions in which the opponent was straight (a) and in which the opponent displayed a gay-stereotypic attribute (i.e., he liked shopping) but was straight (b). High and low internal and external motivation were plotted at 1 standard deviation above and below their respective means. Error bars represent the standard errors of the mean level of aggression within condition.

General Discussion

In support of the plausible deniability hypothesis, identifying an opponent with a gay-stereotypic attribute produced a situation in which the target's stigmatized group status was privately inferred but plausibly deniable to others. Whether or not the situation granted plausible deniability, participants high in internal motivation to respond without prejudice aggressed at low levels and participants low in both internal and external motivation aggressed at high levels. The behavior of participants low in internal but high in external motivation, however, depended on whether the situation granted plausible deniability. Although they aggressed at low levels when the opponent's group membership was explicitly stated, they aggressed at much higher levels when he was described with a gay-stereotypic attribute. Stereotyping to infer group membership grants plausible deniability for prejudice-based aggression, which undermines external—but not internal—motivation to respond without prejudice.

Plausible deniability involves strategically maintaining a nonprejudiced image in the eyes of others, which distinguishes the present work from other work that has emphasized hiding prejudicial responses primarily from the self (e.g., aversive prejudice; Gaertner & Dovidio, 1986). In the current experiment, if plausible deniability made participants themselves unaware of their prejudicial behavior, the participants with high internal motivation, who were personally invested in being nonprejudiced, would have aggressed more when plausible deniability was present than when it was absent. That pattern failed to arise. As such, our findings indicate that plausible deniability neither challenges nor overlaps with theories that emphasize prejudicial behavior that arises in opposition to earnestly held personal values. Rather than being unaware of their bias, people who are low in internal motivation and high in external motivation to respond without prejudice are aware of situational constraints and strategically take advantage of opportunities to express prejudice with impunity (see also Plant & Devine, 1998, 2001, 2009).

These findings have implications for both theory and application, including insights about how we as scientists conceptualize prejudice and how we as a society attempt to reduce prejudice. The present work demonstrates that a publicly known group status is unnecessary to elicit prejudicial responses (see also Blair, Judd, Sadler, & Jenkins, 2002). Moreover, although traditional notions hold that labeling someone as gay may lead to negative outcomes, that label may, counterintuitively, have a protective effect in some instances. Gay or straight people who possess gay-stereotypic characteristics may face higher levels of bullying and violence than gay people who "come out" and are somewhat protected by social condemnations against antigay prejudice (although straight people with gay-stereotypic traits may be unable to reap the benefits of these normative forces). The patterns demonstrated in the present work could play out in many different domains (e.g., hate crimes, bullying in schools, discrimination in the military) and for prejudice against many different groups without visible defining features (e.g., lesbians, Jews, political groups). Similar patterns may also arise for members of groups that do have visible defining features (e.g., racial groups) in situations that allow perpetrators to infer but plausibly deny knowing that the target is a member of a stigmatized group.

When social perceivers use a stereotype to jump to a conclusion about group membership, they are often incorrect, which makes people vulnerable to discrimination or violence on the basis of prejudice against a group to which they do not belong. Indeed, many victims of antigay hate crimes are straight men who were mistakenly assumed to be gay (e.g., Ethington, 2011; Goldstein, 2011; McFadden, 2008; "3 Men Are Charged," 1993; U.S. Department of Justice, 2010). Some policies and laws take these stereotypic inferences into account by condemning discrimination or violence that is based on "actual or perceived" group status (e.g., the Matthew Shepard and James Byrd, Jr., Hate Crimes Prevention Act, 2009). These policies may be difficult to implement, however, because it is challenging or even impossible to demonstrate to an authority (e.g., a court) that someone has made a private stereotypic inference about group membership. Even the prejudice victims themselves may be unsure whether negative behavior toward them is based on group membership, and that attributional ambiguity compounds the negative emotional consequences experienced by the victims (Crocker, Voelkl, Testa, & Major, 1991; Sue et al., 2007).

Institutional antiprejudice policies (i.e., appeals to external motivations) may be effective sometimes for some people, but plausible deniability highlights one way in which normative mandates and policies proscribing prejudice may be insufficient to restrict prejudicial behavior. When responding without prejudice is personally important to people, however, they behave consistently with that value even when they can evade social sanctions against prejudice. To truly reduce prejudice, it is important to make responding without prejudice a personal goal rather than merely a normative mandate (Cox et al., 2012; Devine, 1989; Devine, Forscher, Austin, & Cox, 2012; Plant & Devine, 1998, 2001, 2009).

Conclusion

These findings are both disheartening and encouraging. It is disheartening that despite decades of progress toward eliminating prejudice, merely possessing a stereotypic trait

can make people vulnerable to such high levels of aggression. Nevertheless, it is encouraging that people with personal egalitarian convictions refrain from expressing prejudice even when they have plausible deniability. Contrary to the pessimistic notion that most people avoid expressing prejudice for purely self-presentational reasons, many people refrain from expressing prejudice even when they could express it while maintaining a nonprejudiced public image. Although the present work may have revealed a previously hidden bastion of prejudice, we hope that shining light on plausible deniability will facilitate future research that helps thwart prejudice in all its forms.

Author Contributions

W. T. L. Cox conceived and designed the initial study and refined it with advice from P. G. Devine. W. T. L. Cox supervised data collection, analyzed the data, and wrote the initial draft of the manuscript, which was extensively revised and developed with P. G. Devine. Both authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Lyn Abramson and Jack Dovidio for comments on a previous version of the article. For comments on this article's ideas, we also thank Carlie Allison, Leonard Berkowitz, Markus Brauer, Emily Dix, Patrick Forscher, Sebastian Korb, Chelsea Mitamura, Magda Rychlowska, James Shepperd, Kristin Shutts, and Adrienne Wood, and for their tireless work developing and running this complicated experiment, we thank our experimenters Mathew Baker, Eric Benzel, Lauryn Besasie, Nick Graetz, Renee Kramer, Erica Nagy, Amelia Petermann, Matthew Phillippi, Mark Resnick, Marc Rodriguez, Rebecca Segal, and Gino Tassara.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was funded in part by a University of Wisconsin Kellett Mid-Career Award and a Leon Epstein Faculty Fellowship (both to P. G. Devine).

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. Readers familiar with the prejudice literature may note that plausible deniability has similarities to Gaertner and Dovidio's aversive prejudice construct (Dovidio & Gaertner, 1998; Gaertner & Dovidio, 1986; Hodson, Dovidio, & Gaertner, 2004). These constructs, however, are distinct. Aversive prejudice is specified to be a feature of people whose consciously endorsed values are nonprejudiced, whereas plausible deniability focuses on people who *lack* strong personal motivations against prejudice. Therefore, these people would not be characterized as having the aversive form of prejudice. People who endorse egalitarian values but hide prejudicial behavior from themselves (i.e., those with aversive prejudice) are very different from people who do not personally endorse egalitarianism but are concerned about punishment from others and are therefore consciously strategic about when they express prejudice.

2. As in prior work on aggression, men aggressed at higher levels than women in our experiment. Including participant gender in our analyses does not change the reported patterns of aggression; therefore, we do not discuss gender any further.

3. All data and experimental materials are available publicly at http://www.sciencecox.com/pub/pdshock13.

4. Conducting these analyses using dummy coding and including either or both control conditions as referent groups does not change the patterns of the effects. Because of space limitations, we focus on the most direct tests of our hypotheses. For the full set of more complex analyses, see the Supplemental Material available online.

References

Berkowitz, L. (1989). Laboratory experiments in the study of aggression. In J. Archer & K. Browne (Eds.), *Human aggression: Naturalistic approaches* (pp. 42–61). New York, NY: Routledge.

Blair, I., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*, 5–25. doi:10.1037/0022-3514.83.1.5

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, *37*, 917–934. doi:10.1080/1369183X.2011.576195

Cooper, A. (2011, October 11). The battle to end bullying [Video file]. In *Anderson Cooper 360°*. Atlanta, GA: Cable News Network. Retrieved from http://ac360.blogs.cnn.com/2011/10/11/video-the-battle-to-end-bullying/

Couric, K. (Writer), & Terry, J. (Director). (2012). Bullying: A national epidemic [Television series episode]. In J. Zucker (Executive producer), *Katie*. Burbank, CA: Disney-ABC Domestic. Retrieved from http://www.katiecouric.com/2012/10/02/bullying-a-national-epidemic

Cox, W. T. L., Abramson, L. Y., Devine, P. G., & Hollon, S. D. (2012). Stereotypes, prejudice, and depression: The integrated perspective. *Perspectives on Psychological Science*, *7*, 427–449. doi:10.1177/1745691612455204

Crocker, J., Voelkl, K., Testa, M., & Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology*, *60*, 218–228.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18. doi:10.1037/00223514.56.1.5

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit racial prejudice: A prejudice

habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267–1278. doi:10.1016/j.jesp.2012.06.003

Donnerstein, E., Donnerstein, M., Simon, S., & Ditrichs, R. (1972). Variables in interracial aggression: Anonymity, expected retaliation, and a riot. *Journal of Personality and Social Psychology*, *22*, 236–245. doi:10.1037/h0032597

Dovidio, J. F., & Gaertner, S. L. (1998). On the nature of contemporary prejudice: The causes, consequences, and challenges of aversive racism. In J. Eberhardt & S. T. Fiske (Eds.), *Confronting racism: The problem and the response* (pp. 3–32). Newbury Park, CA: Sage.

Ethington, E. (2011, March 10). Straight GWU student mistaken for a gay man, beaten half to death [Web log post]. Retrieved from http://www.prideinutah.com/?p=10180

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.

Giancola, P. R., & Parrott, D. J. (2008). Further evidence for the validity of the Taylor Aggression Paradigm. *Aggressive Behavior*, *34*, 214–229. doi:10.1002/ab.20235

Goldstein, J. (2011, March 15). Police commissioner calls Queens killing a hate crime. *The New York Times*. Retrieved from http://www.nytimes.com/2011/03/16/nyregion/16hate.html?

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2004). The aversive form of racism. In J. L. Chin (Ed.), *The psychology of prejudice and discrimination: Racism in America* (Vol. 1, pp. 119–136). Westport, CT: Praeger.

Katz-Wise, S. L., & Hyde, J. S. (2012). Victimization experiences of lesbian, gay, and bisexual individuals: A meta-analysis. *Journal of Sex Research*, *49*, 142–167. doi:10.1080/00224499.2011.637247

Matthew Shepard & James Byrd, Jr., Hate Crimes Prevention Act, Pub. L. No. 111-84, §4707, 123 Stat. 2838 (codified in part at 18 U.S.C. §249) (2009).

Matthews, C., & Hill, C. (2011). Gay until proven straight: Perceptions of male interior designers from male practitioner and student perspectives. *Journal of Interior Design*, *36*(3), 15–34. doi:10.1111/j.1939-1668.2011.01060.x

McFadden, R. (2008, December 8). Attack on Ecuadorean brothers investigated as hate crime. *The New York Times*. Retrieved from http://www.nytimes.com/2008/12/09/nyregion/09assault.html

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811–832. doi:10.1037/0022-3514.75.3.811.

Plant, E. A., & Devine, P. G. (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

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of Personality and Social Psychology*, *96*, 640–652. doi:10.1037/ a0012960

Shelp, S. G. (2002). Gaydar. *Journal of Homosexuality*, *44*(1), 1–14.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, *62*, 271–286.

Taylor, S. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, *35*, 297–310. doi:10.1111/j.1467-6494.1967.tb01430.x

3 men are charged with murder in New Orleans anti-gay attack. (1993, November 21). *The New York Times*. Retrieved from http://www.nytimes.com/1993/11/21/us/3-men-are-charged-with-murder-in-new-orleans-anti-gay-attack.html

U.S. Department of Justice, Federal Bureau of Investigation, Criminal Justice Information Services Division, Uniform Crime Reports. (2011). *Hate crime statistics 2011*. Retrieved from http://www.fbi.gov/about-us/cjis/ucr/hate-crime/2011

Zeichner, A., Frey, F., Parrott, D., & Butryn, M. (1999). Measurement of laboratory aggression: A new response-choice paradigm. *Psychological Reports*, *85*, 1229–1237. doi:10.2466/pr0.1999.85.3f.1229