CONCEPTUALIZATION AND EMPIRICAL EXAMINATION OF DIVERSITY TRAINING BACKLASH: THE ROLE OF THE MORAL CREDENTIALING PROCESS

By

Mahl Geum Choi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Human Resources and Labor Relations – Doctor of Philosophy

2023

ABSTRACT

My dissertation seeks to promote the transition of “diversity training backlash” from a general concept that means different things to different researchers to a scientific construct regarding which there is a significant consensus, and to experimentally examine when and how such backlash unfolds in the organizational context. To do so, I conducted a systematic review of the DT backlash literature that critically evaluates existing theorizing and empirical evidence addressing DT backlash. Based on my review, I propose a definition of DT backlash and conceptualize the DT backlash construct by theoretically explicating how it cognitively, affectively, and behaviorally manifests itself. Then, I propose and empirically examine how moral credentialing theory explicates a previously unexamined underlying psychological mechanism of DT backlash. I hypothesized that research participants’ recalling DT-related experiences may morally license trainees before participating in DT, thereby leading to a likelihood of expressing prejudice and discriminatory behavior against minority group members (i.e., increased DT backlash). I also tested how one’s justice perceptions regarding DEI values and the assignment of DT, and individual differences such as social dominance orientation, and belief in a just world moderate the hypothesized relationships. Theoretical and practical implications will be discussed.

ACKNOWLEDGEMENTS

My academic journey has been a transformative experience, not just in terms of intellectual growth but also personal development. I would like to express my deepest gratitude to my advisor, Dr. Mark Roehling, for his invaluable guidance, unwavering support, and scholarly insights throughout my entire journey at Michigan State University. You were not just a mentor for me, but you were and still are my role model. I have not only learned how to become more professional and knowledgeable in this career but also learned how to be humble and open-minded to other people and different ideas. You also showed me the ways to become a good and loving person as a husband, a father, and a grandfather (in a distant future). Thank you.

I extend my heartfelt gratitude to the members of my dissertation committee, Dr. James Dulebohn, Dr. Jason Huang, Dr. Chenwei Liao, and Dr. Frederick Morgeson. Your insightful feedback, constructive criticism, and consistent encouragement have significantly enriched this dissertation. I was very lucky to have you on my dissertation committee, as this experience greatly shaped my identity as a researcher and my vision of what kind of person I want to be in the future. I would like to specifically acknowledge Dr. Chenwei Liao for his exceptional insights and expertise that have been pivotal in refining the empirical examination of my model. Finally, I am truly grateful to Dr. Frederick Morgeson for your advice and support for my dissertation and academic life, which had a significant impact on my growth.

I extend sincere thanks to my parents, Hoo Gon Choi and Sook Hyun Lee, and my parents-in-law, Taeseon Choi and Chunhee Lee, for their unwavering love and support. I have great memories of celebrating important moments together, and those memories motivated me to keep on going. To my sister- and brother-in-law, Daun Choi and Minsung Choi, your warm support was very meaningful throughout my journey.

To my beloved wife, Goun Choi, your love and endless support have been my pillar of strength. I was able to withstand and overcome challenges based on your love and support. In moments of self-doubt, your encouragement and sacrifices were my guiding light. My words cannot express the depth of gratitude I hold, and your love transcends any description. I'm incredibly proud of how we have conquered challenges together. It was not easy, but I firmly believe that it fostered deeper intimacy and mutual understanding between us.

To my dear son, Elliot Han-kiel Choi, you've been a source of pure joy and inspiration throughout this journey, too. Your giggle and innocence were my source of energy. As you grow, I hope you understand that you have been an integral part of this journey, and your impact on my life is truly amazing.

I am indebted to the contributions of every individual who supported and inspired me throughout this journey. I received so much love and support while I was at Michigan State University. I sincerely hope to reciprocate and share what I received to people I will meet and people in the community.

TABLE OF CONTENTS

INTRODUCTION

The importance of addressing workplace diversity issues is widely recognized. The effective management of workplace diversity may increase organizations' ability to attract and retain top talent, contribute to a culture that promotes innovation, and reduce the risk of discrimination that may lead to costly lawsuits and publicity that negatively impact relationships with external stakeholders (Ragins & Ehrhardt, 2021; Richard et al., 2004; Ward et al., 2022). Organizational efforts to address diversity issues commonly include diversity training (DT) to promote employee awareness of diversity issues and develop the core skills needed to work effectively in a diverse workplace (Ragins & Ehrhardt, 2021). Related research demonstrates that DT can increase trainees' knowledge, skills, and reactions toward diversity initiatives (Bezrukova et al., 2016). At the same time, existing research also provides significant evidence that DT can trigger backlash effects that negatively influence training outcomes and result in negative progress toward diversity goals (e.g., Holladay et al., 2003; Holladay & Quinones, 2008; Kidder et al., 2004; Leslie, 2019).

For example, multiple studies found that DT can increase prejudice and negative evaluations toward minority group members rather than decreasing the bias (Heilman, 1994; Heilman & Welle, 2006; Leslie et al., 2014). Also, Dobbin et al. (2015) found that DT programs may only provide benefits for particular race/gender groups (e.g., White women) instead of equally promoting benefits for groups with low representation. Relatedly, DT trainees who strongly identify with 'White' identity or historically advantaged groups may perceive DT as a threat rather than a source to learn (Doosje et al., 1999), and as a result, they tend to attach more strongly to their group (Doosje et al., 2002). Together with these results from academia, the news press and popular media also spotlight that DT can generate "backlash effects," and as a consequence, DT can harm workplace equity and diversity rather than promote it (e.g., Lipman, 2018; McGregor, 2016; Tran, 2021).

Given that DT backlash is a widely recognized phenomenon, and both scholars and practitioners frequently warn of its potential to significantly undermine the effectiveness of DT, the development of DT backlash as a scientific construct has received surprisingly little attention to date. Not infrequently, results of research investigating DT training effectiveness are characterized as involving a “backlash” without any explicit attempt to define or describe “backlash” (e.g., Brannon et al., 2018; Brewis, 2019; Kaplan, 2006). Among those studies that attempt to define DT backlash, there is a lack of consensus regarding how DT backlash should be defined. For example, some researchers argue that DT backlash is a negative emotional reaction (e.g., Lindsay, 1994; Ragins & Erhardt, 2021), whereas other researchers have described DT backlash as involving defensiveness (Ancis & Szymanski, 2001), resistance (Kernahan & Davis, 2007), exaggerated stereotyping (Kalev et al., 2006; Nelson et al., 1996), or negative attitudes and behaviors (Plant & Devine, 2001). More fundamentally, efforts to explicate the DT backlash construct are lacking. That is, there have been only a few, relatively modest attempts to offer a theoretical conceptualization of DT backlash (e.g., Burke & Black, 1997; Bergman & Salter, 2013; Dobbin & Kalev, 2018; Kidder et al., 2004).

A second, related but more specific significant limitation of the existing literature is the lack of attention to the multiple psychological mechanisms by which DT may produce a backlash effect. There is theoretical support for multiple psychological mechanisms playing a role in DT backlash (discussed below). However, based on my review, no existing scholarly work addresses theoretically supportable psychological mechanisms, how they differ from each other, or how DT backlash unfolds over time.

The lack of attention to the definition and theoretical development of the DT backlash construct has impeded the systematic accumulation of empirical evidence pertaining to DT backlash. As a result, although there is a strong consensus that DT backlash occurs and can have a significant negative impact on DT effectiveness, we know little about how DT backlash occurs, why it occurs, or how to reduce it. This dissertation addresses such issues, and the following chapter focuses on the review of the literature and provides the purpose of this dissertation in the beginning.

CHAPTER 1: REVIEW OF DT BACKLASH LITERATURE

Overview

The purpose of my proposed dissertation is to help promote the advancement of the widely accepted concept of “DT backlash” from a general concept to a scientific construct. My dissertation’s effort to achieve that purpose can be divided into four main parts. In Chapter 1, I introduce the first main component of my dissertation. In this part, I conduct a review of research and claims made by scholars relating to the concept of DT backlash, and it draws on multiple theoretical perspectives (e.g., psychological reactance, justice perceptions, and moral credentialing theory). The review will identify and critically evaluate the various current definitions of DT backlash, the limited relevant empirical findings, and significant issues relating to the conceptualization or measurement of “DT backlash.” To my knowledge, it is the first such review.

In Chapter 2, based on the insights gained from the literature, I will theoretically explicate the DT backlash construct and propose a definition of DT backlash that can be operationalized and used to guide future research. I also describe important characteristics of DT backlash, including how DT backlash manifests itself, and who is more likely to experience it.

The third main component is addressed in Chapter 3. Here, my dissertation proposes a study that will involve an investigation of selected theoretical issues. Whereas the focus and theoretical contribution of the first two main components of my dissertation will be broad, because it is not possible in a single study to examine all of the theoretical issues identified and discussed in my dissertation, the study focuses on a narrower set of specific issues. The issues that the study investigates were selected because they relate to what I consider the most unique aspects of my conceptualization of the DT backlash construct. More specifically, the study will empirically assess how DT can lead to backlash via the moral credentialing process and show how the moral credentialing processing of DT can produce adverse training outcomes. I examine how prior *positive* experiences of DT can become a moral credential for trainees, which, in turn, ironically leads to an increase in prejudice and a reduction of DT outcomes. I argue that trainees' positive exposures to DT can have a positive influence on DT reactions and acquiring DT knowledge, yet those may lead to increased prejudice via the moral credentialing process. I also account for how such previous experiences of DT can become either a source of motivation or a reservoir of a moral credential for trainees and examine both the immediate effects and delayed effects of such moral credentialing. I test this idea by applying an experimental design that employs samples from Prolific, one of the online surveying platforms. To maximize moral credentialing effects and eliminate other plausible explanations, I use multiple treatment and control conditions. I also focus on the personality traits of trainees, such as social dominance orientation, and belief in a just world, to test which types of trainees are more likely to engage in such moral licensing processes and DT backlash.

The fourth main component of my dissertation will be the development of a future research agenda for contributing new knowledge on DT backlash, the extent to which it may be problematic, and how it can be effectively managed by organizations in order to better promote and sustain a diverse and inclusive work environment. The agenda will not be limited to the future research implications of the results of my study (the third main component). Rather, it will also include research needs identified in my critical review of the DT backlash research, and for the testing of the relationships suggested in my conceptualization of the DT backlash construct.

Together, the four components of my dissertation will contribute to the DT literature in several significant ways. First, by providing the first critical review of claims and research relating to DT backlash, it will increase awareness of the need for researchers to provide greater attention to how they conceptualize DT backlash.

Second, it will provide a comprehensive definition of DT backlash that: 1) encompasses the various types of negative responses to DT documented in previous studies and practitioner accounts across different disciplines; and, 2) addresses deficiencies in the extent to which current definitions of DT backlash reflect the broader discussions of the DT backlash phenomenon found in the literature.

Third, my dissertation so also responds to the call for researchers to examine when and why DT backlash may occur (Roberson et al., 2017) by: 1) providing a more comprehensive theoretical explication of the multiple psychological mechanisms by which DT may produce a backlash among trainees than what can be currently found in the literature; and 2) conducting a study that empirically examines one such mechanism: moral credentialling. I expect the results of the study to show that having positive previous experiences with DT can produce unintended consequences of DT, suggesting that promoting diversity in organizations through DT involves managing such paradox. The study also offers a unique perspective on trainees' previous exposures to DT. Although DT became one of the most widely and frequently used diversity initiatives (Dobbin et al., 2015), and it is not uncommon for employees to receive multiple exposures to DT, the existing literature seldom discussed how trainees may utilize such previous experiences of DT when they are about to participate in DT. By taking account of the previous DT experiences, my dissertation will provide a rare and unique examination of the effects of recalling DT-related experiences, and how such recalls may impact DT backlash.

Fourth, as previously indicated, my dissertation develops a future research agenda for contributing new knowledge on DT backlash, the extent to which it may be problematic, and how it can be effectively managed by organizations genuinely concerned about promoting and sustaining diversity and inclusion. The identification of critical future research needs is especially important in this area of the literature because of the significant practical relevance of DT training, and the fact that theory and research investigating DT backlash are still at a nascent stage.

In the remaining sections of this chapter, I review the current literature regarding DT backlash literature. My review focuses on critically evaluating current definitions of DT backlash and identifying theoretical perspectives that are used to describe the psychological processes of DT backlash.

What is Diversity Training?

Diversity training (DT) refers to “a set of programs intended to increase individual knowledge, skills, attitudes, and behaviors that can facilitate intergroup relations while reducing trainees’ prejudice and discriminatory behaviors” (Bezrukova et al., 2012, pp. 208), and it serves the organizational goals of creating an inclusive environment for every employee and hiring and retaining employees with underrepresented backgrounds (Devine & Ash, 2022; Bezrukova et al., 2016). This is different from traditional training, which focuses more on improving one’s skills and expertise that may be directly related to enhancing job performance (e.g., Kraiger et al., 1993).

In addition to such differences in goals, DT can be different from traditional training in several ways. First, because DT frequently involves increasing self-awareness about diversity issues and promoting changes in biased attitudes and behaviors that trainees may hold toward other minority groups (Hanover & Cellar, 1998; Law, 1998; Probst, 2003), DT could trigger various emotions, especially negative ones such as anger and frustration (Alderfer et al., 1992;

Burke & Black, 1997). Moreover, promoting diversity in the organization can signal to employees that political correctness is valued in the organization. Such signals may bring one's formerly formed political attitudes about other diversity-related practices, such as affirmative action, that may lead to heated debates (Paluck, 2006). Furthermore, realistic group conflict theory explains that when individuals perceive that intergroup relations are centered around competition for material or symbolic resources, they are more likely to show hostility toward minority group members (Campbell, 1965; Levine & Campbell, 1972). Thus, such a trend that trainees in DT reject the ideas to promote DEI values and become more hostile toward DT and minority group members is not a new phenomenon in the organizations.

Second, unlike other types of training, DT can be applied multiple times throughout one's career starting from undergraduate years (Green, 2000). Although the repetitive nature of DT has become widely accepted, surprisingly little knowledge has been generated regarding questions such as what are the effects of multiple exposures to DT on DT outcomes and DT backlash, and how previous episodes of DT are related to forming pretraining attitudes toward the subsequent DT. Recently, researchers have identified that people may experience diversity fatigue, which is a psychological state of reduced motivation and capacity to learn from and deal with diversity efforts even if they believe in diversity values (Smith et al., 2021), possibly due to multiple exposures to diversity efforts. This further suggests that trainees' pretraining conditions, such as DT fatigue, can determine one's level of DT outcomes. As such, acknowledging the pretraining conditions of DT that are related to trainees' repetitive exposure to DT will generate critical knowledge regarding how managers should approach when implementing DT.

Method of Review

For the purpose of this review, I conducted searches on research databases (e.g., Proquest, EBSCOhost) and collected research publications such as journal articles, books, and conference papers), using the search terms: “diversity,” “inclusive,” coupled with the “training,” “initiative,” “program,” “workshop,” “backlash,” “backfire,” and “unintended.” The search returned over 300 publications, and then I further identified research publications by going through the references of the above publications.

Current Definitions of DT Backlash

Although numerous articles and studies use the term backlash or backfire related to DT, there was no consensus on the definition, and some studies even used the term without providing a definition (e.g., Chavez & Weisinger, 2008; Dover et al., 2019). Among those researchers who provide a definition, there is significant variation in the definitions. Table 1 in Appendix A summarizes the current definitions of DT backlash. The following observations and insights are based on a systematic evaluation of the current DT backlash definitions. First, while all definitions describe DT backlash as involving some types of adverse responses to DT training by trainees, some definitions include a broad range of responses (e.g., Burke & Black, 1997; Lee, 2022), and other definitions focus more narrowly on specific types of responses (e.g., emotional reactions; Lindsay, 1994; Ragins & Erhardt, 2021). For example, Lee (2022) recently defined workplace backlash as “the explicit/implicit, and/or intentional/unintentional attempts to reject efforts to promote diversity, taken by both dominant and subordinate social group members to maintain the group-based social hierarchy at work”, while Ragins and Erhardt (2021) focus narrowly on negative emotional reactions that trainees feel during and after DT. Second, the definitions vary in the extent to which backlash responses are described as involving DT outcomes that are less positive than the goal of training versus training that was actually antithetical to the goals of the DT (i.e., outcomes that made the situation worse in terms of achieving diversity goals). That is, while some definitions indicate that backlash only occurs when the response is "antithetical to diversity goals" (e.g., Brannon et al., 2018) or involves "negative progress" toward such goals (e.g., Leslie, 2019), other definitions appear to characterize responses among some trainees that are merely "less favorable" as involving DT backlash (Kidder et al., 2004). The latter approach would characterize a response that makes DT less effective for some training as involving backlash even if the response was not antithetical or counterproductive to diversity goals. Third, although most definitions do not address who may experience backlash, some definitions appear to define DT backlash as a phenomenon that occurs among specific groups (e.g., high-status group members; Burke & Black, 1997; Kidder et al., 2004). Fourth, no definition explicitly identified the level at which DT backlash occurs. Most definitions appear to view DT backlash as an individual level phenomenon. However, definitions that focus on, or at least emphasize, DT backlash among certain groups (e.g., Whites) could be viewed as indicating that DT backlash is a group-level phenomenon.

Psychological Mechanisms of DT Backlash

A number of authors have provided empirical and anecdotal evidence of psychological mechanisms why DT may backfire. In this section, I review the psychological mechanisms of DT backlash that have been identified or theorized by former studies. Table 2 in Appendix B summarizes the psychological mechanisms of DT backlash. Among DT backlash mechanisms, negative emotions are commonly reported by researchers. For example, DT often employs an instruction that may arouse shame, blame, and guilt among trainees, and it may highlight the differences among groups by making trainees too sensitive about the differences, leading to increased tension among groups (Beaver, 1995). In addition, the contents of DT, which is about creating changes and breaking the status quo, may activate negative emotions of dominant group members, such as anger, frustration, and guilt (Lindsey et al., 2020). More specifically, when socially disengaging emotions or negative discrete emotions are aroused during DT, including anger and frustration, it can further reduce the learning outcomes and transfer outcomes of DT. On the other hand, the marginalized group members may also experience shame and embarrassment during DT as some content or instructors may ask them to openly share their experiences as a minority with other trainees (Anand & Winters, 2008; Paluck, 2006). According to the intergroup contact theory, emotions are a key mechanism that can reduce intergroup prejudices (Miller et al., 2004; Voci & Hewstone, 2003), suggesting that eliciting negative emotions during DT may produce a backlash effect. Thus, such experiences of negative emotions during DT can become a basis for trainees' resistance toward the subsequent DT.

Furthermore, trainees may perceive DT as constraining their freedom, and such perceptions may trigger trainees' psychological reactance, which refers to unpleasant motivational arousal caused by perceptions of reduced personal freedom (Brehm, 1966). For example, by confronting white males during training in order to make them realize and acknowledge some biases that they may hold (Burke & Black, 1997), such strategies not only elicit negative responses but also may generate perceptions that DT or the instructor is constraining autonomous motives for self-regulation and autonomy (Beaver, 1995; Brannon, 2018). Trainees who experience psychological reactance due to reduced freedom may restore such motivational arousal by expressing greater prejudice (Plant & Devine, 1998, 2001).

Dobbin and Kalev (2018) further explain why DT may fail to reap intended or positive outcomes by explaining that DT may activate trainees' stereotypes and facilitate more prejudice rather than reducing it. More importantly, along with Dobbin and Kalev (2018), Leslie (2019) also identifies that when organizations offer DT for their employees and strong endorsement for DEI initiatives, such support may make employees believe that their organization is bias free and blind about their own discriminatory behavior. Such an illusory sense of fairness toward the organization may further provide dominant group members with the sense that the status quo is natural, or discrimination does not exist within the organization (Kaiser et al., 2013). Kaiser and colleagues (2013) argue that when such an illusory sense is pervasive in the organization, the members will become more hostile toward the claims that minority group members are experiencing discrimination in the organization. This suggests the group dynamics between dominant and subordinate groups may differently affect each group member's perceptions or bias toward DT.

Relatedly, facilitating DEI initiatives and offering DT within the organization may provoke unfairness perceptions among trainees. The majority group members might interpret DT and diversity initiatives as a signal of reverse discrimination that minority groups are gaining more advantages and, therefore, receiving preferential treatment (Bergman & Salter, 2013; Black & Burke, 1997; Leslie, 2019). This may further lead to a belief or perceived threat that the dominant groups are about to lose their status, resources, and opportunities to be successful within the organization, thereby perceiving DT as a potential that may further lead to reverse discrimination (Bergman & Salter, 2013). Thus, DT backlash may occur because it involves unfairness perceptions based on the existing group dynamics between the majority and minority groups within the organizations. And, more importantly, such group dynamics and relationships between groups may determine one's attitudes (e.g., fairness perceptions) before joining DT (Alderfer, 1992; Paluck, 2006), which, in turn, may trigger DT backlash.

Finally, the moral credentialing or moral licensing theory can explicate why DT backlash occurs in a paradoxical way. The moral credentialing theory describes that people’s good deeds and suppression of prejudice at Time 1 can provide moral credentials that may increase the likelihood of engaging in morally questionable behavior or expressing greater prejudice at Time 2 (Monin & Miller, 2001). Previous researchers have introduced the idea that such a human tendency might be closely related to backlash against diversity initiatives and DT (e.g., Dobbin & Kalev, 2018; Leslie, 2019). Leslie (2019) argues that when an organization sends signals that it highly values morality, employees may use such signals to morally license themselves (e.g., “because my company is valuing morality, I may be moral too”). As a result, they will be less likely to monitor their behavior and may show more subtle discrimination against minority group members. In a similar vein, because participating in DT has moral values, for some trainees, mere participation in DT may be sufficient to self-license themselves. As such, they might be more likely to express prejudice or discriminatory behavior after DT. This idea will be formally hypothesized and empirically examined in Chapter 3.

In the following chapter, I conceptualize the DT backlash construct based on insights gained from the critical review of the literature in Chapter 1. It incorporates and expands upon the previous, relatively modest current efforts to conceptualize DT backlash (e.g., Burke & Black, 1997; Kidder et al., 2004; Leslie, 2019; Sanchez & Medkik, 2004). I start by defining a DT backlash construct to address the current limitations in the current definitions and theoretically explicate different manifestations of, and other important characteristics of DT backlash.

CHAPTER 2: CONCEPTUALIZATION OF DT BACKLASH

Definition of the DT Backlash Construct

I define DT backlash as an individual trainee's experience of resistant or hostile responses against DT or minority group members as a consequence of taking DT that manifests itself in the forms of cognitive, affective, and behavioral responses that are antithetical (counterproductive) to achieving DT goals. In this definition, minority group status can be based on categories based on gender, race, age, religiosity, and sexual orientation, and some members may hold multiple minority categories. Additionally, DT goals in my definition refer to the previously mentioned goals of DT, which are about promoting intergroup relations and reducing prejudice and discrimination (Bezrukova et al., 2016). My decision to conceptualize and define DT backlash as a phenomenon that occurs at the individual level reflects the consensus view of the academic literature. Although previous definitions of DT backlash do not explicitly address the level at which it is viewed as occurring, and some definitions seem to suggest DT backlash may be a group level phenomenon (e.g., Lindsay, 1994), most discussions of DT backlash describe an individual level phenomenon. In addition, when studied empirically, DT backlash has consistently been operationalized at the individual level (e.g., Holladay & Quinones, 2008; Kidder et al., 2004). I recognize that there may be circumstances in which a group of employees may share a similar backlash response to DT, and researchers may be interested in investigating DT backlash at the group level in those circumstances. However, my focus on individual level responses in defining and investigating DT backlash is supported by the consensus approach to DT backlash found in both the academic and practitioner literature.

In contrast to some definitions that focus on a specific manifestation of DT backlash (e.g., emotional reactions, Ragins & Erhardt, 2021), my proposed definition of DT backlash incorporates the wide range of manifestations of DT backlash that have been identified and supported in the literature. The manifestations of DT backlash include proximal responses, such as immediate reactions to DT, to distal responses, such as making negative progress in promoting diversity by bolstering one's prejudice or by expressing biased attitudes toward other people or the organization. Trainee's immediate reactions to DT, such as trainees' feelings toward DT and instructor (Holladay & Quinones, 2005; Rynes & Rosen, 1995), are often the crucial indicators of DT that can readily determine whether DT was successful (Wentling & Palma-Rivas, 2000) because training reactions, in general, are closely associated with subsequent learning and transfer outcomes (Alliger et al., 1997; Giangreco et al., 2010; Sitzmann et al., 2008. Additionally, DT backlash may involve negative cognitive responses, such as intensified stereotypes and resistance to acquiring more diversity knowledge during and after taking DT. Thus, when trainees experience DT backlash, their learning outcomes and transfer outcomes of DT will be diminished. On the other hand, DT can produce relatively distal responses regarding one's affective changes, and behavioral changes, which involve dissipating training motivation and self-efficacy (affective responses), strengthening one's explicit and implicit prejudice (attitudinal responses), and based on these responses, trainees will be more likely to engage in discriminatory behavior.

A critical aspect of the proposed definition is its specification that DT backlash does not merely involve resistant, negative, or adverse responses to DT. Rather, the negative or adverse response must be antithetical (counterproductive) to diversity goals. That is, backlash occurs only when the DT has the opposite of the desired effect (e.g., it increases the trainee's negative attitude toward diversity or their expression of prejudice). This aspect of the definition also suggests that DT that elicits trainees' proximal and/or distal DT backlash may prevent the overall effectiveness of DT because both responses of DT backlash will negatively influence training outcomes. This conceptualization is consistent with the lay concept of "backlash." The Merriam-Webster Dictionary defines backlash as "a sudden violent backward movement." In addition, the specification that the negative or adverse response must be antithetical to diversity goals contributes to the discriminant validity of the DT backlash construct by distinguishing DT backlash from broader considerations of DT training effectiveness. For example, DT that simply has no effect or is simply more effective for some trainees than others does not meet the proposed definition of DT backlash.

DT Backlash versus DT Ineffectiveness

There is merit in further clarifying the relationship between DT backlash and DT ineffectiveness because the two constructs share conceptual similarities. I argue that DT backlash is one of the special forms of DT ineffectiveness. Both DT backlash and DT ineffectiveness are negative outcomes of DT, and they should negatively influence enhancing DEI values. As a counterpart to DT effectiveness, DT ineffectiveness is something that can impede effective learning or training transfer. DT ineffectiveness can broadly capture negative reactions and outcomes after one's participation in DT, and it also can include other negative goal progress in DEI values, such as DT backlash. It may further include stagnancy in learning, inattentiveness during DT, less commitment toward DEI values, and dissatisfaction with DT. These training outcomes may result in unproductive learning and the status quo in achieving DEI goals. These outcomes are likely to negatively influence DEI values like DT backlash, but only in indirect and less obvious ways than DT backlash. I contend that these types of DT ineffectiveness do not necessarily increase the risks of aggravating DEI progress. On the other hand, DT backlash involves manifestations that are directly counterproductive to reaching DEI goals, and it involves more obvious manifestations such as increased prejudice and discriminatory behavior. As these characteristics suggest, not all forms of DT ineffectiveness will provoke negative goal progress toward DEI values, whereas DT backlash will harm DEI values in more direct and discernable ways.

I further contend that, as one of the specific forms of DT ineffectiveness, DT backlash could occur even when most parts of DT were successfully achieved. Although observing training outcomes impeded by DT backlash could be more common, some trainees might be able to cognitively learn the materials, while feeling anger towards DT or minority group members. This can happen not only because DT effectiveness and DT backlash focus on different aspects of training outcomes, but also because they have fundamentally different psychological mechanisms. When such mechanisms are activated during training, trainees may experience a DT backlash regardless of the effectiveness of DT. Because DT backlash deals with negatively-valenced outcomes such as anger, hostility, or explicit prejudice, it is often omitted in evaluating DT effectiveness. However, if researchers or practitioners fail to include measures assessing these manifestations of DT backlash when evaluating DT effectiveness (or ineffectiveness), they may fail to recognize that DT backlash is occurring, and as a result, be deprived of the opportunity to address it.

Manifestations of DT backlash

DT backlash may take several different forms (Jackson, 1999; Kidder et al., 2004). For example, after participating in DT, some trainees might perceive that information from DT was unfair, some may feel anger and guilt, and others may become more prejudiced against the minority group members. As such, there are three primary dimensions of DT backlash, and it is closely associated with various criteria of DT: 1) cognitive responses, 2) affective responses, including emotional and attitudinal responses, and 3) behavioral responses. I demonstrate how each dimension of DT backlash emerges and how researchers can capture such negative responses using such dimensions. Thus, in this part, I review the existing DT backlash studies and categorize the various manifestations of DT backlash according to widely recognized criteria for assessing training effectiveness.

1. Cognitive responses to DT backlash

One of the widely measured criteria in DT is cognitive learning or diversity knowledge related to gaining more information and acquiring knowledge about the values of diversity and DEI initiatives (Bezrukova et al., 2016; Lindsey et al., 2020). With regard to training in general, cognitive learning includes improving one's knowledge, skills, and abilities (KSAs) in targeted areas (e.g., Kirkpatrick, 1959), or it can include verbal knowledge, knowledge organization, and cognitive strategies to learn more about the subject matter (Kraiger et al., 1993). In the domain of diversity training, researchers agree that cognitive outcomes involve gaining more knowledge or information, and the content of knowledge is often more related to prejudice or biases that trigger discrimination or organizational policies regarding DEI issues (Lindsey et al., 2020). For example, Kulik and Roberson (2008) summarize that cognitive learning involves gaining specific knowledge about organizational policies or resources related to diversity issues.

Because DT cognitive outcomes are about gaining knowledge, researchers, in general, measure them by administering knowledge tests after DT. Cognitive outcomes of DT are often reported as the most effective, consistent outcome generated by DT, and some studies describe that an increase in diversity knowledge tends to be maintained over time while attitudinal or behavioral learning decreases after some time (Bezrukova et al., 2016; Kulik & Roberson, 2008). A recent meta-analysis also supports these trends of diversity knowledge acquired during DT as the cognitive outcomes of DT had the second strongest effect size following the reactions to DT (Bezrukova et al., 2016). Such a result, however, does not eliminate the possibility that DT backlash occurs in this domain.

To conceptualize DT backlash in cognitive dimensions, I draw on Kraiger et al. (1993)'s definition of cognitive outcomes in general training literature and integrate stereotypical knowledge structure to such cognitive outcomes. From Kraiger et al. (1993)'s perspective, there are three dimensions of cognitive outcomes: declarative knowledge (i.e., acquisition of knowledge), knowledge structures (i.e., organization of knowledge elements), and metacognitive skills (i.e., mental strategies that manage knowledge acquisition or application). According to these authors, cognitive learning outcomes not only involve obtaining knowledge and information but also include the complex processes of synthesizing existing knowledge elements and developing strategies to acquire new knowledge. Psychologists and organizational behaviorists agree that cognitive learning outcomes are not just about attaining more declarative knowledge (Bell et al., 2017; Chung et al., 2022; Day et al., 2001; Ford et al., 1998; Lacerenza et al., 2017; Stanhope et al., 2013). Scholars also noted that declarative knowledge is a necessary but not sufficient condition in achieving higher-order knowledge development (Ackerman, 1987, Anderson, 1982). Thus, successful learning outcomes of DT should represent gains in declarative knowledge, an effective organization of such knowledge, and the development of learning strategies in DEI values and goals.

Researchers suggest that experts and novices are not only different in the breadth and depth of declarative knowledge but also vastly differ in knowledge structures and metacognitive activity. For example, experts tend to develop complex knowledge structures or mental models that help them to interpret the situation and events while seeking optimal solutions (Chi et al., 1989). In addition, experts tend to possess higher levels of metacognitive skills, such as assessing the difficulty of new knowledge or being aware of their performance during the tasks (Pressley et al., 1987).

In the context of DT, DT provides trainees the opportunity to acquire knowledge regarding psychological mechanisms of discrimination (e.g., prejudice, stereotypes), DEI policies, issues, and ways to promote DEI values in the workplace (Bezrukova et al., 2016). When DT functions well, trainees will not only focus on obtaining more declarative knowledge, but also they will learn how to organize their declarative knowledge elements and build cognitive maps of how such elements are interrelated with each other. Also, such trainees may develop their metacognitive skills that involve planning, monitoring, and changing behaviors toward learning objectives (Brown et al., 1983) in order to facilitate effective learning about DEI values and issues.

I contend that when DT backlash occurs in the cognitive dimensions, it will particularly affect knowledge structures and metacognitive skills rather than influencing all three dimensions. At its face value, my definition of DT backlash is about the trainees losing or unlearning what they already know about DEI values, policies, and issues (i.e., declarative knowledge). However, my definition of DT backlash in this dimension is not to argue that trainees will remove existing knowledge structures in their cognitive systems, but it is more about trainees becoming resistant to building any knowledge structures or metacognitive skills during DT. More importantly, previous researchers found that individuals are also capable of developing stereotypical knowledge structures and changing their existing knowledge structures to become stereotypical (Garcia-Marques et al., 2006).

Given such theoretical foundation and findings, I argue that when cognitive manifestations of DT backlash occurs, trainees 1) will be more likely to modify their existing knowledge structures to incorporate more stereotypical information and knowledge, 2) will be less likely to utilize their metacognitive skills or activities during DT, resulting in the poor acquisition of knowledge regarding DT, and 3) will become more stereotypical.

2. Affective responses of DT backlash

Reducing affective responses of DT backlash is highly critical in achieving the goals of DT as those are related to making trainees realize their own prejudice and reducing discriminatory behaviors based on such learning. Affective responses can encompass a broad range of responses to DT backlash as, in general, they may include attitudinal and motivational outcomes from training (Kraiger et al., 2013). I maintain that there are three components included in the affective manifestation of DT backlash: Negative emotions, affective responses, and attitudinal responses. Negative emotions are one of the affective manifestations of DT backlash that researchers and practitioners have frequently reported, and those emotions involve various negative emotions, such as anger, shame, and frustration (e.g., Anand & Winters, 2008; Jackson, 1999; Paluck, 2006). According to my definition of DT backlash, negative emotions represent immediate and proximal responses to DT backlash, and they can affect other dimensions of DT backlash, which will eventually harm the effectiveness of DT. This process is similar to the traditional training model that demonstrates behavioral changes and performance improvement are affected by reaction to training (Noe, 1986; Noe & Schmitt, 1986)

In addition to negative emotions, attitudes are generic and broad determinants of behavior (Allport, 1929, 1935), and changing trainees' prejudiced attitudes toward DEI initiatives can be a key mechanism in reducing trainees' discriminatory behavior and enhancing goals toward achieving DEI initiatives. In traditional training literature, attitudes are often measured as training satisfaction, how trainees felt about the training and contents, and it often includes training motivation and self-efficacy (Kirkpatrick, 1994; Towler, 2003). In diversity training literature, however, attitudinal criteria, such as general attitudes toward diversity and DEI initiatives, and specific prejudice or attitudes toward different minority groups, are targeted learning outcomes (Bezrukova et al., 2012; Kulik & Roberson, 2008; Lindsey et al., 2020). Additionally, some specific types of DT are geared toward improving trainees' awareness of DEI initiatives as well as their biases and prejudice rather than facilitating behavioral changes based on role modeling and active practicing (Bezrukova et al., 2012; Kulik & Roberson, 2008). In terms of DT backlash, the affective/attitudinal dimension can be potentially the most important dimension because it is the dimension where trainees' prejudice is located, and such prejudice is one of the primary drivers of overt discriminatory behavior or subtle discrimination (Kulik & Roberson, 2008). Thus, failure to reduce this dimension will transfer its negative effects to the behavioral dimension of DT backlash, increasing the likelihood of discriminatory behavior.

To provide a clearer conceptualization of affective/attitudinal dimensions, I make a distinction between affective and attitudinal dimensions of DT backlash, although the two are, in general, very similar and often paired together (e.g., Bezrukova et al., 2016). I categorize an affective dimension as training attitudes that are related to oneself, including but not limited to justice perceptions, psychological reactance, training self-efficacy, and training satisfaction. On the other hand, I classify the attitudinal dimension of DT backlash as solely associated with one's attitudes toward minority groups. By doing so, I highlight DT backlash attitudes' inward (the affective dimension) versus outward (the attitudinal dimension) perspectives. Affective DT backlash may be more common to observe than in other dimensions because the meta-analytic findings (Bezrukova et al., 2016) indicate that affective outcomes of DT were more difficult to improve while those outcomes were also more difficult to be retained over time.

Another vital aspect of affective responses of DT backlash is that this manifestation could be conceived as either processes or outcomes of DT backlash because it can trigger behavior. That is, this dimension could be seen as an outcome of DT backlash as one's level of prejudice increases and one's perceptions of injustice increase, but, at the same time, affective/attitudinal responses may further enact one's discriminatory behavior or other behavior related to DT backlash; in this way, affective/attitudinal responses could be viewed as processes that lead to behaviors.

i. Emotional responses of DT backlash

Various negative emotions can be triggered by participating in DT. Because trainees will develop their attitudes toward DEI initiatives and DT before joining DT, DT involves more politically and emotionally charged learning processes (Alderfer, 1992; Jackson, 1999; Paluck, 2006). Numerous researchers have also contended that trainees, especially majority group members, may induce negative feelings against DT, minority groups, or facilitators. Negative discrete emotions, such as anger (e.g., Anand & Winters, 2008), guilt (e.g., Kalev et al., 2006; Kowal et al., 2013), and anxiety (Burke & Black, 1997) have been identified as common discrete emotions that trainees may feel after taking DT. When historically advantaged groups were confronted with or accused of their biases, such members showed more feelings of anger and contempt (Pendry et al., 2007; Mollica, 2003). In other cases, anger and contempt can be elicited during DT in which materials or instructions of DT blatantly blame the majority group members or threaten the integrity of their groups (Branscombe & Wann, 1994; Mackie et al., 2000). On the other hand, minority group members can also feel shame and embarrassment when they are pinpointed to share their experiences as minority group members representing the whole group (Anand & Winters, 2008; Jackson, 1999). Also, because DT is related to initiating changes in the organization, it arouses uncertainty in trainees, and, as a result, trainees may feel anxious about the consequences of the changes (Burke & Black, 1997). Furthermore, when instructions and contents of DT are not clear, trainees may feel that they just have to be careful around minority group members (Anand & Winters, 2008).

Negative emotions from DT backlash include feeling the aforementioned negative emotions, but these emotions could be too broad to form one's resistance toward DT. Trainees who feel negative emotions may end up learning something during DT. For example, when trainees feel guilt, they may transfer such negative emotions to further motivate themselves in order to learn more and reduce prejudice. Thus, it is critical to further distinguish emotions that are core to DT backlash and that can account for triggering one's resistance and hostility against DT. Lindsey and colleagues (2020) have argued that socially disengaging emotions, which highlight separateness from other people (Kitayama et al., 2000, 2006) and focus on increasing social distance with others (Lazarus, 1991), are the primary emotions driving DT backlash. Unlike socially engaging emotions (e.g., empathy, guilt, and shame) that are associated with valuing interdependence, social harmony, and accepting relational embeddedness (Kitayama et al., 2006), socially disengaging emotions, such as anger, pride, and frustration, are low in social orientation and associated with maximizing personal interests in exchange of social harmony (Kitayama et al., 2000; Lazarus, 1991). Socially disengaging emotions could be either positive (e.g., pride, feeling superior) or negative in valence (e.g., anger or frustration).

In the DT context, where sharing and improving understanding of differences to facilitate social harmony is its ultimate goal, socially engaging emotions, such as state empathy, have been conceptualized as a key explanatory mechanism that can improve trainees' altruistic motivation to change, which in turn will enhance DT outcomes (Lindsey et al., 2020). For socially disengaging emotions, there has been surprisingly no empirical evidence or frameworks that explain the relationship between DT backlash and socially disengaging emotions. It is reasonable to expect socially disengaging emotions to be associated with reduced social orientation (i.e., creating harmony), feeling such emotions may increase trainees' prejudice and tendency to engage in discriminatory behavior (Lindsey et al., 2020). For example, people feeling pride after taking DT might lead to similar processes of enhanced prejudice and discriminatory behavior because feeling pride, which is a discrete emotion that highlights social distance from oneself to others (Lazarus, 1991). Consequently, by feeling proud after taking DT, such trainees may focus more on their ability to gain knowledge and the certificates that they acquired for taking DT and may focus less on transferring their knowledge to behavior to create social harmony. Hence, feeling proud during or after DT might increase the likelihood of DT backlash. According to Kitayama and colleagues (2006), anger is also socially disengaging emotion that is less socially oriented, and it is often triggered when the person's goal pursuit is disturbed by certain obstacles. If DT trainees perceive DT as something unnecessary that is preventing their tasks or performance, taking DT should make trainees angry, and this may lead to other manifestations of DT backlash.

Taken together, negative emotions constitute immediate and relatively short-lived responses to DT. Because negative emotions are important proximal outcomes that can facilitate other manifestations of DT backlash, they are related to increased prejudice, which lies in the attitudinal dimension of DT backlash, and increased discriminatory behavior.

ii. Affective responses of DT backlash

Negative affective responses against DT can be one of the most prevalent forms of DT backlash. Previous studies have shown that taking DT often elicits negative affective responses, such as psychological reactance against DT (Brannon et al., 2018), perceived unfairness (Kidder et al., 2004), and a zero-sum perspective on DT (Bergman & Salter, 2013). Psychological reactance, justice perception (or perceived unfairness), and the zero-sum perspective of DT are the primary mechanisms that occur in affective dimensions. Psychological reactance refers to one's unpleasant motivational arousal triggered by experiencing intimidation or feeling of losing their personal, behavioral freedom (Brehm, 1966). In general, people experience psychological reactance when they are obligated to learn particular perspectives that are against their beliefs (Brehm & Brehm, 1981), and such tendency is intensified when DT is enforced or mandated by organizations (Graupmann et al., 2012; Rosenberg & Siegel, 2016). Also, studies from the field of communication indicate that persuasive messages often trigger reactance that is expressed via counterarguing and anger (Dillard & Shen, 2005; Kim et al., 2013, Rains, 2013). In addition, when such messages use controlling and coercive language, individuals experience more reactance and perceive them as more intimidating (Miller et al., 2007; Quick & Stephenson, 2008). In line with political conservatism's emphasis on personal freedom, some people tend to reject the ideas of political correctness and DEI initiatives because they believe promoting such ideas is constraining others' freedom (Paluck, 2006). Because DT is often geared toward reducing one's prejudice by correcting misconceptions and misdeeds (Bezrukova et al., 2012), some trainees perceive the contents and messages of DT as constraints of personal freedom to speak or express their thoughts freely (Dobbin & Kalev, 2018). For example, researchers have argued that DT may make the groups not targeted by DT (i.e., white males) feel guilty or become sensitive about the group differences (Karp & Sammour, 2000), triggering whites' psychological reactance.

According to this theory, when people experience the unpleasant motivational state of reactance, they attempt to restore their freedom by utilizing cognitive and behavioral efforts that are often accompanied by negative emotions (Steindl et al., 2015). On a cognitive side, a person who is experiencing reactance tends to diminish the value or attractiveness of the source of threat or escalate the perceived freedom (Brehm, 1966; Brehm & Brehm, 1981; Bushman & Stack, 1996). On the behavioral side, intimidated individuals attempt to restore their freedom by directly showing restricted behavior or engaging in aggressive behavior to vent out their negative emotions, such as anger (Brehm & Brehm, 1981; Dillard & Shen, 2005; Rains, 2013). In this sense, when some trainees experience reactance against DT, to balance their reactance, they will be more likely to augment their prejudice and perform discriminatory behavior and less likely to accept the DT contents and correct their misconception.

In addition to psychological reactance, some trainees often perceive the contents of DT as unfair because they tend to see it as preferential treatment for the minority group members in exchange for decreasing the chances of success for the majority group members (Leslie, 2019). As members of a dominant group, some white males perceive DT as an organizational vehicle of an unfair intervention that can prevent them from preserving their current status, while increasing the chance of success for minority group members (Leslie, 2019; Mobley & Payne, 1992). As such, after taking DT, participants may experience a heightened level of unfairness perceptions, especially for the majority group members. Likewise, when trainees perceive DT contents or DT itself is harming distributive or procedural justice, there will be a greater likelihood to elicit hostility and resistance against DT that can even further impact other dimensions of DT backlash. Kidder and colleagues (2004) examined how employees' recent promotion (or not getting a promotion) affects backlash against DEI initiatives and found that when employees did not receive a promotion, they were more likely to react negatively to the initiatives. Kidder et al. (2004) conclude that such an effect was due to lower distributive justice resulting from not getting promotions. In a similar vein, when employees perceive that procedures in implementing affirmative action programs are unfair, they are more likely to contest the programs (Bobocel et al., 1998). In terms of DT, when trainees believe that they were unfairly assigned to DT, such as by other colleagues recommending the focal trainee to be in DT, they tended not to acquire diversity knowledge and, further, they were more likely to engage in preferential treatments for certain racioethnic groups (Sanchez & Medkik, 2004). Hence, when trainees perceive DT as procedurally unfair during DT due to its contents or the way they were assigned to DT, trainees will be more likely to restore such unfairness by bolstering or expressing their prejudice.

Finally, DT can elicit one's zero-sum perspective that demonstrates how whites often see practices or policies facilitating equality as losses to their groups and gains for the marginalized groups (Eibach & Keegan, 2006; Norton & Sommers, 2011). In their experimental study, Eibach and Keegan (2006) found that people who were told to hold zero-sum perspectives and consider that they are losing their share due to DEI practices tend to perceive such practices as making greater progress toward racial equality compared to the groups that were told to maintain the gain perspective. Furthermore, survey-based research (Norton & Sommers, 2011) reports that whites were more likely to believe racism is based on a zero-sum game and view decreasing prejudice against Blacks causes increasing prejudice against whites. The zero-sum game perspective against DT is also related to falsely believing reverse discrimination is more concerning than discrimination against minority groups (Bergman & Salter, 2013).

Through these three psychological mechanisms, trainees may find more room to hold onto and express their previous prejudice and biases rather than to recognize and concentrate on mitigating such false beliefs.

iii. Attitudinal Responses

The attitudinal dimension of DT backlash is directly associated with the goal of DT and DEI initiatives as DT intends to reduce trainees' biases. Yet, it is known as one of the difficult dimensions to observe improvements and maintain the positive effects of DT over time (Bezrukova et al., 2016). Empirical and anecdotal evidence from DT literature specifies that trainees' level of prejudice against minority groups is often increased rather than decreased (e.g., Hood et al., 2001; Neville & Furlong, 1994; Stewart et al., 2003). I categorize the attitudinal dimension of DT backlash as the dimension associated with explicit and implicit prejudice against other groups. Prejudice refers to "a negative evaluation of a social group or a negative evaluation of an individual that is significantly based on the individual's group membership" (Crandall & Eshleman, 2003, pp. 414). Because prejudice involves conscious psychological processes, one may justify or suppress prejudice in a given context (Crandall & Eshleman, 2003).

Reducing explicit prejudice against other minority groups is one of the DT's objectives, but both qualitative and quantitative reviews of DT often report that DT's effectiveness in mitigating prejudice is somewhat limited (Bezrukova et al., 2016; Kulik & Roberson, 2008). Such explicit prejudice measures directly ask respondents about either their blatant forms of prejudice against various racioethnic groups (often called "old-fashioned" forms of prejudice) or more subtle types of prejudice, such as the Modern Racism Scale (McConahay, 1986), which asks the extent to which the respondents agree with subtle statements about privileges and attitudes toward blacks. Because respondents may answer such direct measures of explicit prejudice with extreme levels of social desirability, researchers are becoming more concerned about using such explicit measures of prejudice and are utilizing the test that measures one's implicit bias (Lindsey et al., 2020).

Likewise, Dovidio and Gaertner (1998, 2000, 2004) report that whites' overt expression of prejudice against blacks has been decreasing at the national level, but they argue that aversive racism (Gaertner & Dovidio, 1986), which is subtle and rationalizable discrimination against minority performed by whites who believe in egalitarian values. Utilizing the implicit association tests (IAT; Greenwald et al., 1998), which measure the response latencies to stereotypic category paring (e.g., whites paired with good) using a computer, may further assist researchers in addressing the social desirability of explicit prejudice and aversive racism. For example, when individuals are faster to pair "blacks" with something "bad" than pairing "whites" with those bad traits, the test will conclude that the individuals have an automatic preference for whites over blacks.

Some scholars have argued that DT may result in stereotype activation (Dobbin & Kalev, 2018). Stereotype activation refers to increased accessibility to a distinctive set of characteristics that are believed to be associated with social categories (Wheeler & Petty, 2001). Based on the stereotype activation theory (Wheeler & Petty, 2001), researchers have proposed that DT can activate stereotypes inherent in groups by making the group membership more salient and highlighting differences between each racial group (Anand & Winters, 2008; Beaver, 1995; Duguid & Thomas-Hunt, 2015; Egan & Bendick, 2008; Heilman, 1994; Sidanius et al., 2001). Furthermore, when people are asked to suppress their stereotypes, people, ironically, tend to become more prejudiced as their stereotypes become consciously accessible (Dobbin & Kalev, 2018; Galinsky & Moskowitz, 2000; Wegner, 1989, 1992), and researchers have discussed that similar processes occur for DT (e.g., Dobbin & Kalev, 2018). For example, when people were made to suppress their stereotypes against skinheads, it actually led to a higher level of stereotypes and more discriminatory behavior toward skinheads (Macrae et al., 1994). Researchers have further revealed that both expressions and suppressions of stereotypes activate subsequent stereotypes (Liberman & Forster, 2000), and suppression tends to trigger the accessibility of stereotypes and counterstereotypes (Galinsky & Moskowitz, 2007). As a result, some of the majority group members (nontargets of DT) may conclude that the minority group members are people who need help and are incompetent, thereby bolstering their prejudices (Leslie, 2019). In the context of DT backlash, such post-suppressional rebounds of stereotypes could be a primary driver for bolstering one's implicit prejudices.

Existing intergroup relationships within the organization could also be associated with attitudinal responses of DT backlash. Based on realistic group conflict theory, Brief et al. (2005) demonstrate that an increase in actual diversity in organizations was associated with decreased job satisfaction and increased perceived conflict. This result suggests that organizational efforts to increase diversity in organizations can threaten some employees, leading to DT backlash. Relatedly, social identity theory and social categorization theory explain that employees may engage in in-group/out-group categorization and competition based on existing group differences within the organization in order to clearly delineate their group identity (Brewer & Brown, 1998; Tajfel & Turner, 1979). In other words, people have a tendency to favor ingroups over outgroups, and to strengthen the in-group identity, they also tend to categorize the in-group and out-group. In doing so, they tend to focus on group differences, which often strengthen their deep-seated prejudice against the out-group. And, researchers have found empirical evidence that such a tendency increases one’s stereotypes and prejudices (Hugenberg & Bodenhausen, 2004; Perdue et al., 1990; Tajfel & Forgas, 2000; Van Knippenberg & Dijksterhuis, 2000).

3. Behavioral responses of DT backlash

In general, the behavioral responses of DT backlash are also the dimension that is directly related to achieving the primary goal of DT because if the trainees do not change their behavior, there is less possibility of seeing systematic changes toward DEI initiatives within the organization (Devine & Ash, 2022). Along with the affective/attitudinal responses of DT backlash, failure to stop backlash in this dimension may further indicate that DT has actually produced unintended negative outcomes rather than yielding intended outcomes. Also, the enactment of the behavioral dimension could be the culmination of other dimensions regarding DT backlash (i.e., cognitive, affective responses) because the behavior may reflect the failure in such dimensions.

The behavioral criteria of DT and the behavioral dimension of DT backlash focus on similar sets of behaviors. Both may include measures of behavioral intentions or behaviors that are related to promoting/preventing DEI initiatives, discriminatory behavior against historically minor groups, and behaviors that affect the representativeness of minority groups. For instance, these include behaviors that are aligned/misaligned with DEI initiatives that often discourage discriminatory behaviors (Hanover & Cellar, 1998), and decisions to hire or promote minority group members (Kulik et al., 2000). Both of the dimensions may include the expressions or regulation of prejudice toward minority group members, and one’s willingness to engage in or actual behavior toward minority groups, such as providing support or help for those groups.

Among many theoretical explanations, the moral credentialing theory has received attention from multiple scholars (e.g., Kalev & Dobbin, 2008), but only a few researchers have established the link between moral credentialing and DT backlash (e.g., Leslie, 2019). Moral credentialing or moral licensing describes the phenomena in which a person's positive behavior (e.g., helping others) can provide a moral credential for the person, and with such a credential, the person is more likely to perform morally questionable or immoral behavior in a subsequent situation (Merritt et al., 2010; Monin & Miller, 2001; Miller & Effron, 2010). Research regarding moral credentialing theory shows that acquiring moral credentials is positively associated with expressing prejudice toward minority groups and less willingness to hire or support minority group members (e.g., Bradley-Geist et al., 2010; Mann & Kawakami, 2012; Monin & Miller, 2001). Relating this theory to DT backlash, I argue that some trainees will morally license their participation in DT, which may further lead to the expression of prejudice against minority group members or less willingness to hire minority group members, which will result in the underrepresentation of minority groups in the organization. In other words, previous experiences of DT could become a source of moral credentials for trainees, and it may lead to enhanced prejudice and discriminatory behavior in subsequent situations.

Paradoxically, such tendencies may be strengthened if the organization emphasizes DEI initiatives because individuals are known to vicariously license the moral deeds of other entities by observing the moral behavior and having membership (e.g., other persons and organizations; Ahmad et al., 2020; Kouchaki, 2010; Leslie, 2019). In addition, Yam and colleagues (2017) found that when people engaged in a good deed with external motivation, it increased their psychological entitlement via moral credentialing, further leading to workplace deviance. This occurred because individuals who performed good deeds based on external motives felt they were under-rewarded and licensed such good deeds to balance their feeling of getting fewer rewards. Taken together, empirical results suggest that when an organization facilitates DEI initiatives and when trainees have higher levels of external motivation (e.g., coerced to participate in DT), it might increase the likelihood of morally credentialling of DT because people may attempt to compensate themselves by licensing the experience of DT to balance their feeling of getting underpaid. As such, in an organization that emphasizes the values of diversity, trainees may rely on those positive aspects of the organization to obtain moral credentials (e.g., "I am not biased because my organization has strength in promoting diversity"), and this tendency may be intensified when DT is externally motivated (e.g., "I am not biased because my organization has strength in promoting diversity and participating in DT is enough"). Trainees with this type of moral credentials may not monitor their behaviors or attitudes, thereby increasing the expression of prejudice and discriminatory behavior (Leslie, 2019). This shows the complex processes of DT backlash that involve multiple dimensions and mechanisms in triggering such unintended negative outcomes for DT.

Trainee Characteristics and DT Backlash

Thus far, not many studies have theorized how trainee characteristics may influence DT backlash or DT effectiveness. However, a recent conceptual piece (Roberson et al., 2022) introduces the idea that trainees' personas may pose different challenges in achieving effective learning and transfer outcomes. These personas are introduced in this section because these individual differences could also elicit DT backlash. According to these researchers, trainees can be categorized as possessing defensive, anxious, and overconfident personas (Roberson, 2022). Defensive trainees are individuals who tend not to believe in DEI values or the necessity of DT, and they tend to view DT as a symbolic gesture. Anxious trainees tend to consider DT as a threat to their current employment status or something that leads to their awkward behaviors or mistakes. Also, they are more likely to experience guilt and shame during DT. Overconfident trainees are described as viewing DT as a chance to demonstrate what they know rather than to learn more about DEI content. These trainees believe that they are cofacilitators or cotrainers of DT.

I argue that the trainee personas are likely to predict the likelihood of experiencing DT backlash because it is a unique form of DT ineffectiveness. With their suspicious view toward the goals of DT and DEI values, defensive trainees could be more likely to feel anger and hostility against DT. Anxious trainees, who worry that DEI values will erode their current status or their stereotypes might come out during DT, may overly consider group differences and their own prejudice, making their biases more salient during DT, which may strengthen one's prejudice. Lastly, overconfident trainees could be more likely to self-validate themselves with previous DT experiences and knowledge that they obtained. The moral licensing theory explains that when individuals morally validate themselves, they are more likely to express prejudice in the subsequent situation, suggesting that DT backlash may occur. Overall, these personas suggest that trainees' individual differences may play a vital role in regulating their emotions or biases, thereby leading to DT backlash.

Who is more likely to experience DT Backlash?

As mentioned in the previous part, there is far more evidence about extant DT backlash studies have focused on white males as a group that may mainly experience DT backlash because they tend to possess higher status and because DT challenges the status hierarchy within the organization (Burke & Black, 1997; Kulik & Roberson, 2008). As a result, the messages of DT can be seen as a potential threat to white males and/or as preferentitreatmentnts for minority group members (Leslie, 2019, Kalev & Dobbin, 2008). Furthermore, DT with a narrow focus tends to make white males feel left out of the inclusion that DT is attempting to promote, and, as a result, white males become more hostile and angrier to the diversity values and DT (Holladay et al., 2003, 2008).

Although the DT literature does not often highlight the backlash of minority groups, they might experience DT backlash as well. Minority group trainees may become embarrassed and disappointed in DT and DEI initiatives if they perceive that they are being used as tokens within the organization (Brannon et al., 2008; Jackson, 1999). According to Jackson (1999), trainees from ethnically minor groups report that their experiences and feelings are often misunderstood by instructors. As a result, DT often elicits anger and frustration from minority group members when the instructors fail to deliver the topic with empathy and sensitivity. Also, when DT focuses too much on intergroup differentiation and intergroup conflicts, both dominant and minority group members may feel negative emotions, thereby leading to DT backlash (Anand & Winters, 2008). More specifically, the dominant group members may feel ashamed or angry because they may be accused of transgressors and feel fearful or anxious because DT often only provides vague and abstract guidelines on how to act or behave. On the other hand, similar to the phenomenon of social identity threat (Steele et al., 2002), minority group members may feel anxious when they are unsure of how the dominant groups will react to DT (Pietri et al., 2019). For example, in their experimental study, Pietri et al. (2019) found that the gender diversity interventions that attempted to reduce gender biases among men and women actually led to lower levels of women’s sense of belonging in their expert areas due to heightened social identity threat. Additionally, for some minority group members (targets of DT), participating in DT may become an unpleasant occasion in which they confirm that their coworkers are more prejudiced than they previously believed (Anand & Winters, 2008).

Targets of DT Backlash

Extant literature on DT backlash suggests that trainees experiencing backlash could target different entities to behaviorally enact their hostility. In general, common targets of DT backlash are the racioethnically minority groups, DEI initiatives, and the organization (Burke & Black, 1997; Leslie, 2019; Mobley & Payne, 1992). When DT backlash comes out as interpersonal aggression against minority group members, trainees will tend to perform discriminatory behavior, including verbal and physical aggression toward the minority group members, but it can also take a similar form to counterproductive work behavior toward individuals (CWB-I), which includes acting rudely and withholding information (Fox & Spector, 1999; Robinson & Bennett, 1995). As previous research regarding discriminatory behavior against minority groups and workplace aggression has noted, such behaviors can take overt or subtle forms (Jones et al., 2016). Overt behaviors that target the marginalized groups may include verbal and physical harassment, and intentional forms of discriminatory behavior, while subtle forms of discriminatory behavior may involve intentional ostracism and degrading of the ideas or identities of the minority groups (Thomas & Plaut, 2008).

When DT backlash overtly enacts itself toward DEI initiatives or the organization that endorses such initiatives, it can be expressed via forms of vandalism, inappropriate graffiti, or counterproductive work behavior toward the organization (CWB-O; e.g., taking excessive breaks, intentionally working slow, and stealing; Robinson & Bennett, 1995; Spector et al., 2006; Yang & Diefendorff, 2009). When experiencing DT backlash, trainees may engage in CWB-O to restore their experiences of perceived distributive and procedural injustice (Berry et al., 2007; Hershcovis et al., 2007). Subtle forms of backlash behaviors that aim to harm the organization may include reducing organizational citizenship behavior toward the organization (OCB-O) and avoiding DEI initiatives or inequities. Not always intentional, but by engaging in the subtle forms of DT backlash, trainees erode DEI initiatives and values.

In summary, although there are areas of agreement in describing the DT backlash phenomenon, there is no consensus definition of the DT backlash construct. Moreover, all existing definitions are limited in one or more ways in the extent to which they capture the DT backlash phenomenon researchers and scholars have expressed interest in, or the ability to provide guidance regarding the operationalization of the DT backlash construct (discussed further below). Finally, previous research fails to reflect or address various manifestations of DT backlash using a single definition.

Pretraining Conditions and Training Design Features Affecting DT Backlash

Research suggests that pretraining conditions may also be a factor that can significantly affect the attitudinal dimension of DT backlash. In traditional training literature, researchers have emphasized and found that trainees' pretraining expectations and motivation are critical predictors of training effectiveness (Tannenbaum et al., 1991), and pretraining contexts (e.g., training assignment) that can affect one's pretraining characteristics, such as justice perceptions, self-efficacy, and motivation to learn, further predicted training outcomes (Quinones, 1995). Focusing on DT, Hanover and Cellar (1998) found that trainees' perceptions about the extent to which the DT aligns their personal values formed by close friends, family members, and popular media was associated with pretraining attitudes, which, in turn, influenced the post-training outcomes.

Here, I introduce some of the identified DT characteristics and pretraining conditions that can impact the attitudinal dimension of DT backlash. When trainees were assigned to different motivational conditions (autonomous vs. external/controlled), trainees in the external motivation groups showed a higher level of both explicit and implicit prejudice compared to trainees in the autonomous motivation condition (Legault et al., 2011). Also, Sanchez and Medkik (2004) showed that when trainees felt unfair about their assignment to DT, it increased the level of prejudice. This suggests that trainees' motivation previous to joining DT and DT characteristics, such as whether DT is offered in a mandatory way, can affect the level of DT backlash that they might experience. Although many studies report that mandatory DT has a positive effect on DT effectiveness, results from the primary studies suggest that mandatory DT can also elicit DT backlash for some trainees via external motivation.

The focus or content of DT may also affect the attitudinal manifestation of DT backlash. Narrowly-focused DT, such as training focusing only on race or gender compared to focusing comprehensively on other categories, gender, disability, age, and sexual orientations, tends to make people believe that the organization does not consider the complexity of diversity issues (Holladay et al., 2003). Narrowly-focused content of DT often make the trainees, especially white men, perceive that they are left out of the inclusion and DEI initiatives and may further make such trainees feel guilt or hypersensitive, leading to DT backlash (Brannon et al., 2008; Burke & Black, 1997; Karp & Sammour, 2000). In a similar vein, such contents can also make the group differences more salient and make attitudes toward other groups more extreme. The narrowly-focused DT tends to foster the zero-sum perspective regarding DT (Brannon et al., 2008).

Moreover, training features such as communicating strategies or approaches can greatly impact DT backlash. Previous studies found that when diversity practices or policies use a color-blinded approach, which emphasizes highlighting the sameness among groups and ignoring the racial differences, in delivering its contents, individuals demonstrated increased prejudice (e.g., Apfelbaum et al., 2008; Norton et al., 2006). On the other hand, when the DT is communicated using a multiculturalism approach that focuses on valuing differences among groups, it reduces prejudice and leads to more collaboration in performing tasks (Correll et al., 2008; Norton et al., 2006). Some researchers suggested that using the multiculturalism approach for DT may greatly reduce the backlash effects, and the dominant group members may become more likely to understand DT's goals and motivations, leading to training success (Brannon et al., 2008; Rynes & Rosen, 1995).

Feedback Loops of DT

As noted briefly in the previous section, the current literature on DT and DT backlash does not fully address how the outcomes of DT may feed back into trainees’ pretraining attitudes or motivation. In their process-oriented model of DT effectiveness, Roberson et al. (2022) conceptualize that there are three stages regarding DT, which are pre-training, training, and post-training stages. Consistent with the traditional training literature (e.g., Salas et al., 2012), this model proposes that each stage requires specific factors to produce positive learning outcomes from DT (Roberson et al., 2022). The pre-training stage is where managing and framing trainee’s cognitive, motivational, and affective readiness toward DT becomes imperative. In the training stage, training characteristics such as training contents and group composition during DT are influential factors for DT effectiveness. For the post-training stage, factors that enhance training transfer, such as supervisor and coworker support, positively influence how trainees change attitudes and behaviors based on learning. The model suggests that DT backlash in one stage will negatively impact the subsequent stage.

Although it is critical to understand DT effectiveness as a process, it is also important to recognize that such a process could be cyclical in nature. A majority of studies in the DT context typically treated DT as a single unique episode, but I contend that DT backlash should carry over to the subsequent DT, impacting the pre-training readiness and attitudes. For example, increased hostility and resistance during the training stage may not only seriously impede one's knowledge gain and exacerbate DT transfer after DT, but also such attitudes may remain until the pre-training stage of the next DT. That is, DT backlash experience in an earlier episode of DT will determine one's level of DT readiness before participating in the subsequent DT. Other researchers have also noted that trainees join DT with differences in motivation and attitudes (Holladay & Quinones, 2008; Paluck, 2006). Such negative pre-training attitudes will create DT backlash in the training and post-training stages, completing a vicious cycle of DT backlash.

In the following chapter, I empirically test one of the psychological mechanisms of DT backlash, a moral credentialing process, using an experimental design. According to moral credentialing theory, a person's good deeds at Time 1 can offer moral credentials for the person, which, in turn, makes the person more likely to conduct morally questionable or immoral behavior at Time 2. Applying this rationale to DT backlash, I examine how trainees' previous DT experiences may offer moral credentials to trainees, leading to greater bias and discriminatory behavior. The method and samples will be discussed.

CHAPTER 3: DISSERTATION STUDIES EXAMINING THE ROLE OF THE MORAL CREDENTIALING PROCESS IN DT BACKLASH

Theoretical perspectives that were reviewed in the previous chapters, such as psychological reactance and stereotype activation, help describe and explain how DT can contribute to employees' overall resistance to DEI initiatives and elicit a backlash among trainees. However, as others have observed, multiple theoretical perspectives are needed to fully explain why and how backlash occurs (Dobbin & Kalev, 2018; Leslie, 2019), and other theoretical perspectives suggest that DT backlash could involve much more intricate psychological mechanisms that have been the focus of researchers' attention to date. To help provide a more complete picture of how and why DT backlash occurs, in this chapter, I draw on moral licensing theory (or moral licensing theory) to develop hypotheses regarding how DT can become a source of moral credentials, which, in turn, will lead to subsequent expression of prejudice. I then describe and report the results of two experimental studies that test my hypotheses. More specifically, the two studies experimentally examine how pretraining conditions of trainees (e.g., recalling the number of previous DT and positiveness of DT) may trigger moral credentialing processes and how such processes will be amplified among trainees who perceive receiving DT is unfair before coming to DT, thereby leading to a higher level of DT backlash.

THEORETICAL BACKGROUNDS & HYPOTHESES DEVELOPMENT

Moral credentialing theory (Monin & Miller, 2001) posits that when individuals acquire moral credentials by engaging in good behavior or suppressing prejudice, they are inclined to balance such good deeds by subsequently conducting bad deeds (e.g., Yam et al., 2017) or expressing prejudice (e.g., Monin & Miller, 2001). For example, when people showed disagreement with blatant racist remarks, the same people were more likely to express prejudice in the following situations (Monin & Miller, 2001). Such human tendencies are likely to occur in the context of DT because mere participation in DT can become a license for some trainees. I argue that when people perceive DT and DEI initiatives as unfair practices, they may instrumentally use their experiences of DT participation as a moral certificate to express their prejudice or discriminatory behavior, which are the behavioral responses of DT backlash. This idea is in line with the findings that when people are externally motivated to do good deeds, those good deeds are more likely to become a moral license (Yam et al., 2017). I further examine whether such moral licensing processes also affect trainees' levels of explicit and implicit prejudice against minority group members. Generating in-depth knowledge about such complex mechanisms will shed light on how to reduce DT backlash and how to better facilitate DEI initiatives through DT.

Additionally, although it is highly likely that employees have experienced multiple DT throughout their careers as the values of DT and DEI initiatives became more important for organizations (Roberson et al., 2001), the prior studies have not adequately accounted for how one's prior experiences of DT, and one's perceptions of the organization's DEI policies may influence the following DT. I argue that trainees' positive experiences of DT can negatively affect DT backlash through moral licensing processes. That is, trainees' positive experiences or perceptions of DT may provide a moral licensing for some trainees that will result in DT backlash (Dobbin & Kalev, 2018; Leslie, 2019). Not all trainees, however, will undergo moral licensing processes because it is likely that the positive experiences of DT will motivate trainees to learn from DT and facilitate DEI initiatives. To tease out the moral licensing effects and motivating effects of positive pretraining conditions, I contend that trainees who believe DT is an unfair practice will be more likely to experience moral licensing processes instead of motivational processes. More specifically, I argue that trainees who perceive DT as deteriorating the fair distribution of rewards between dominant and subordinate groups or as procedurally unfair to their group status will be more likely to use their prior experiences of DT as moral credentials to justify their prejudice and discriminatory behavior. Furthermore, to shed light on the role of individual differences in these complex relationships, I examine the three-way interactions between pretraining conditions, justice perceptions of DT, and personality traits such as social dominance orientation (SDO) and belief in a just world (BJW). By testing the three-way interactions, I highlight who is more likely to experience DT backlash via moral licensing processes.

Moral Credentialing Theory & DT Backlash

Moral credentialing theory (Monin & Miller, 2001) posits that when people engage in moral behavior or successfully control their prejudice to emerge, these behaviors are likely to strengthen their moral identities, which, in turn, increase the likelihood for people to perform morally questionable behavior and express prejudice in a subsequent situation. That is, acting in a socially desirable direction in one situation allows actors to conduct the opposite behavior in the subsequent situation (Mullen & Monin, 2016). Monin and Miller (2001) found that disagreeing with a discriminating statement at Time 1 licenses the subjects to engage in behavior that is opposite from their previous behavior at Time 2 (i.e., hiring a white person over a black person when they have the same resume). Furthermore, follow-up studies found that people are more likely to morally license themselves and express willingness to hire a white candidate for the job after disagreeing with the racist statement (Effron et al., 2012) and expressing support for Barack Obama (Effron et al., 2009). Moreover, when people recalled their past good behaviors (e.g., Blanken et al., 2015; Jordan et al., 2011) or their positive personality traits (e.g., Blanken et al., 2015), it licensed people to express reduced intentions to help others. In other domains, researchers found that people who chose green products over regular products were more likely to license themselves, leading to increased cheating on the following tasks (Mazar & Zhong, 2010). Importantly, a meta-analysis unveiled that moral credentialing effects were largely significant across multiple studies, though it yielded a small-to-medium effect size (average Cohen's $d = 0.31$; Blanken et al., 2015).

In the organizational contexts, researchers have found support for the ideas that performing organizational citizenship behavior (OCB) or helping others can morally license the actors, leading to subsequent counterproductive work behavior (Klotz & Bolino, 2013) and workplace deviance (Yam et al., 2017). Another line of research found that moral licensing effects occur between leaders and followers, and one study (Lin et al., 2016) found evidence that followers' ethical behaviors provide moral credentials for the leaders, who will become more abusive to their followers in the subsequent interactions. Similarly, Ahmad et al.'s (2020) results indicate that by observing followers' helping behavior, the leaders were more likely to use follower's helping as a source of moral credentials, which eventually led to leaders' unethical behavior. Taken together, there is significant support for the proposition that moral credentialing not only occurs when the actors engage in socially desirable behavior themselves, but it may also result from observing related people's behavior (Effron & Monin, 2010). That is, what is referred to as *vicarious* moral licensing may occur (Kouchacki, 2011). This can be highly relevant for DT backlash contexts as participating in DT and endorsing DEI initiatives are becoming more socially desirable and moral for the organizations, and employees may vicariously license themselves from such positive organizational actions (Leslie, 2019). Also, in the DT context, the focal trainee’s vicarious moral licensing may occur by observing peers’ engagement toward DT, even though this goes beyond the scope of this study.

To date, surprisingly little attention has been given to the role that moral credentialing might be expected to play in DT backlash. Based on my review, although several researchers have mentioned that moral credentialing may play a role in DT backlash (e.g., Bohnet, 2016; Dobbin & Kalev, 2018), only Leslie (2019) provides a significant discussion of that potential role. In describing DEI initiatives that “backfire,” Leslie (2019) reasons that DEI initiatives signal various messages to employees, and a particular signal, the message that ethical values are emphasized by the organization, can actually make them believe that their organization is free of bias. Such perceptions can serve as moral credentials, which, in turn, trigger subtle discrimination against the targets of DEI initiatives. Building on this work, this study suggests that the backlash processes of DT involve more complex psychological processes than one might consider.

The extant literature also gives surprisingly little attention to the effect that pretraining conditions may have on DT backlash. The general training literature identifies and emphasizes that the pretraining motivation and conditions of trainees are significantly associated with training outcomes (e.g., Ilgen et al., 1979; Quinones, 1995). However, there has been a paucity of attention on pretraining conditions in DT literature. Given that trainees take multiple DT per year and during their time in college education (Green, 2000), a lack of knowledge in this area is preventing the researchers and practitioners from getting a descriptive picture of who comes to DT, with what in their minds, and how it can influence DT outcomes and DT backlash. Pretraining conditions for DT are often shaped by trainees’ beliefs about DEI initiatives or diversity practices such as affirmative action (Alderfer et al., 2012; Paluck, 2006). Training characteristics that can influence one's pretraining attitudes and motivation, such as titles of DT (Holladay et al., 2003; c.f., Ratner & Miller, 2001), whether DT is mandatory or not (Kaplan, 2006), and one's prior exposures to DT (Roberson et al., 2001) have been found to influence DT outcomes and DT backlash. In their empirical study, Holladay and colleagues (2003) found that when the title of DT was more straightforward ("Diversity Training") compared to the comprehensive, encompassing title ("Valuing Differences" or "Working Together"), trainees came to DT with more expectations to experience backlash from DT.

In this study, I identify and examine three critical pretraining conditions that can elicit one's moral credentials, further leading to DT backlash, and those are: 1) recalling a previous experience of DT, 2) recalling the number of previous DT, and 3) recalling the positive aspects of the current organization's DEI initiatives. Because taking DT already conveys positive, ethical values, recalling the previous DT might make those experiences salient enough for the moral credentialing process to occur. Previous research has identified that even briefly recalling their previous moral behaviors often licenses individuals to reduce their willingness to help others and donate to others (Blanken et al., 2015; Jordan et al., 2011). Furthermore, an experimental study found that recalling positive experiences with a black individual has led to more willingness to hire a White person over a Black person for the job (Bradley-Geist et al., 2010). Such evidence suggests that when trainees recollect their satisfied memories of previous DT, it may be easier for trainees to license their previous DT experience, which may lead to more chances to express prejudice in the subsequent DT.

It is also possible that recalling the number of previous DT may produce moral licensing effects. Because participating in DT is socially desirable and moral, recalling the number of previous DT can become a good source of moral credentials for trainees. This is because recalling the number of previous DT can evoke relatively tangible and technical information from previous experiences rather than inducing people's specific experiences of DT. That is, recalling the number of previous DT could function as punch cards that some trainees may find easier to morally license and become more entitled to their previous DT. Thus, I posit that trainees who recall more numbers of DT may be more likely to self-license before participating in DT, further leading to DT backlash after DT.

More importantly, the moral licensing theory describes two paths through which moral licensing processes can occur. One is through acquiring moral credentials, and the other is through obtaining moral credits. The former processes explain that performing good deeds provides moral credentials that are like a certificate that bulletproofs their subsequent bad deeds (Miller & Effron, 2010). The latter processes are similar to balancing moral bank accounts using good deeds and bad deeds as a currency (Nisan, 1991). In this sense, recalling one's satisfying experiences with previous DT should be more closely related to moral credentialing processes, whereas recalling the number of DT can elicit information regarding their current moral credits. These mechanisms will be discussed in more detail in the later section.

According to Leslie (2019), organizational signals that indicate ethics are strongly valued in the organization can be a source of moral credentials for employees, which, in turn, may lead to backfire effects or subtle discrimination. This occurs because employees may believe that their organizations are free of prejudices, which will be likely to become the object of moral credentials, leading to less self-monitoring and self-regulation. Moreover, some researchers found that people tend to acquire moral credentials by observing others' moral deeds (i.e., vicarious moral licensing; Ahmad et al., 2020; Kouchaki, 2011). In their study, Ahmad and colleagues (2020) found that leaders often take credit for their followers' OCB and self-license themselves in conducting unethical leadership behavior. More importantly, Kouchaki (2011) revealed that when the actors received information about one of the group members' previous nondiscriminatory behavior, it led them to conduct prejudiced behavior in subsequent situations. In the DT setting, when trainees believe that their organizations provide strong and healthy DEI initiatives, they may be more likely to take credit for being a member of such an ethical organization. This is in line with social identity theory (Hogg & Terry, 2000; Tajfel & Turner, 1986), such that organizational members may become more receptive to group norms and tend to depersonalize when organizational contexts are salient.

Moral Credits vs. Moral Credentials

As mentioned above, previous research has identified that there are two distinct psychological mechanisms in the moral licensing process: one is through moral credits, and another is through moral credentials (Miller & Effron, 2010). The former psychological mechanism, moral credits, describes that the moral licensing process is similar to managing a balance on moral bank accounts. Based on the idea that an individual's moral self-concept is dynamic and changes with an individual's moral actions toward moral equilibrium (Nisan, 1991), the moral credits model proposes that one's moral deed provides a person with moral credits, which can be spent for the subsequent immoral deed. By engaging in such unethical behavior, the actors achieve moral equilibrium, which is often the desired state according to the moral credits model (Jordan et al., 2010; Zhong et al., 2009). On the other hand, when the actors gain moral debts by performing immoral deeds, then they may attempt to balance their moral accounts by doing moral deeds subsequently (i.e., moral cleansing; Tetlock et al., 2000; Zhong & Liljenquist, 2006). One notable feature of this perspective is that when the actors perform bad behavior to balance out their previous good behavior, the actors tend to know they are about to engage in immoral behavior while knowing that doing so would be socially acceptable because both the actors and observers recognize that the actors have excessive moral credits to spend (Jordan et al., 2010; Zhong et al., 2009). In other words, the actors acknowledge that their subsequent behavior would be and appear to be immoral.

On the other hand, based on the study of Monin and Miller (2001), the second psychological mechanism, which refers to the moral credential model, highlights that the moral credentialing process involves how individuals change construing their subsequent immoral deeds after the initial moral deeds (Miller & Effron, 2010). Unlike the moral credit model, this model explains that when the actors conduct the subsequent bad behavior, they tend to view such behavior as morally justifiable and, thus, may not recognize that their behavior would be immoral. That is, the moral credential model posits that doing moral deeds in Time 1 tends to make them think their subsequent behavior in Time 2 would be within the moral boundary. From this perspective, it can be said that acquiring moral credentials provides the actors with distorted views of the subsequent immoral behavior. Due to such distortion in their construal, the actors are able to preserve their moral self-concept intact even though they perform immoral behavior. Additionally, the moral credits model suggests that the actors have to conduct multiple good deeds to conduct the multiple bad deeds, whereas the moral credentials model theorizes that the actors can revisit the past good deeds multiple times to license themselves (Zhong et al., 2009).

I test both models in this part. I argue that recalling previous experiences of DT and their organization’s strengths in DEI initiatives will elicit moral credentialing paths as these pretraining conditions are likely to provide credentials than credits by focusing on overall perceptions of their experience. In other words, by thinking about the previous DT experiences and the positiveness of DEI initiatives in the organization, trainees may be able to reacquire multiple moral licenses from psychologically visiting the same previous experiences (or from the overall experiences of DT) rather than gaining one moral credit from the overall experience. This idea highlights the moral distortion of the actors and the slogan for the moral credentialing process that the actors are becoming “entitled to transgress” (Merritt et al., 2010, pp. 348). However, recalling the number of DT taken will highlight the moral credits paths by making their previous experiences as countable experiences. By engaging in this process, the trainees may demonstrate increased expressions of prejudice while knowing that they may be transgressing moral boundaries. Based on these rationales, I formally hypothesize the following:

Hypothesis 1a: Recalling previous experiences of DT will increase DT backlash

Hypothesis 1b: Recalling the number of DT that participants have taken will increase DT backlash

Hypothesis 1c: Recalling the positive aspects of DEI policies at the current workplace will increase DT backlash.

Moral Credentialing Effects vs. Consistency Effects

Of importance, prior researchers demonstrate that positive behavior at Time 1 does not always license people to engage in unethical behavior at Time 2. That is, a good deed that people engage in may produce licensing effects and lead to subsequent bad behavior, whereas the very same good deed may motivate people to engage in good behavior in the subsequent situation. The latter refers to consistency effects because people are conducting the same or similar behavior from Time 1 to Time 2. (e.g., Blanken et al., 2015; Effron & Conway, 2015; Miller & Effron, 2010; Mullen & Monin, 2016). Such consistency effects are aligned with a stream of research highlighting how previous behavior leads to more of the same behavior (e.g., Baumeister et al., 1994). For example, Baumeister and colleagues (1994) found that when people fail to reach a goal, they tend to feel guilty about it and strive more in the subsequent attempt. Teasing out licensing effects from consistency effects is particularly important for this study because it is equally possible to predict that pretraining conditions of DT can lead to motivation effects instead of moral licensing effects. Furthermore, not every person with the same pretraining conditions of DT will experience moral licensing and DT backlash. Thus, to eliminate or find other plausible explanations and shed light on when and how DT backlash occurs from pretraining conditions of DT, identifying and testing the moderators that elicit moral credentials for the trainees is critical.

Researchers have identified moderators that further predict when and how good deeds will lead to subsequent good or bad deeds (e.g., Conway & Peetz, 2012; Effron et al., 2009). In particular, research suggests that people tend to pursue self-consistency rather than moral credentialing when they focus on abstract goals of their behavior because such an abstract construal provides individuals opportunities to think about superordinate goals and moral values (Conway & Peetz, 2012). Furthermore, when people have low identification or low commitment toward a goal, they are more likely to license their actions toward the goal, preventing themselves from reaching the goal (Effron et al., 2009; Koo & Fishbach, 2008; Mullen & Monin, 2016). On the other hand, individuals with a high level of identification showed less tendency to license their progress and were more motivated to reach a goal, bolstering their initial behavior. This occurs because the low-committed actors may think they have made enough progress toward the goal or demonstrated a token commitment toward the goal, making it easier for them to license their initial behavior (Mullen & Monin, 2016). Taken together, the literature suggests that individuals with low identification or low commitment will be more likely to perceive their progress toward a goal or initial behavior as *sufficient*, and such perceptions will further escalate one’s tendency to license their initial behavior.

In the DT backlash contexts, identifying such moderators would greatly enhance the understanding of when either consistency effects or moral credentialing effects are likely to emerge. For example, people who already have negative attitudes toward DT before participating in DT might be more prone to morally license their previous experiences of DT to express their prejudice against minority group members and DEI initiatives. In a similar vein, when individuals were externally motivated to help other people, those individuals were more likely to feel psychologically entitled, subsequently leading to workplace deviance (Yam et al., 2017).

Moral Credentials and DT Backlash: The Moderating Role of Justice Perceptions

I argue that one’s overall perceptions of distributive and procedural justice regarding DT are crucial factors that can determine one’s level of the moral credentialing process, which will affect DT backlash. Researchers have noted that when trainees perceive that DEI initiatives are only promoting minority groups' status at the expense of their group status, trainees may feel that DEI initiatives as preferential treatment (Leslie, 2019). Additionally, when trainees believe that they are unfairly assigned to DT, they tend to show diminished DT outcomes (Sanchez & Medkik, 2004). As such, when people perceive DEI initiatives or DT as unfair, they would be more likely to join DT with negative attitudes.

More specifically, when trainees believe that the fair chance of receiving rewards is restrained by promoting diversity in the organization and by the organizational implementation of DT, trainees may join DT with negative attitudes against DT and other minority groups. That is, trainees’ perceptions of distributive justice, which focuses on the fairness of outcomes based on their contribution or inputs to outcomes (Adams, 1965), can negatively affect pretraining attitudes before joining DT, which will impact their DT outcomes. Distributive justice is a workplace attitude that is related to job satisfaction, organizational commitment, job performance, and OCB (Colquitt et al., 2001). For example, employees with low distributive justice are inclined to be less satisfied with their jobs, less effective at their work, and significantly more likely to withdraw from their work (Colquitt et al., 2001; Lind & Tyler, 1988). Because the perceptions of the number of outcomes received and contributions that people make are based on subjective evaluations (Adams, 1965), employees may hold different levels of distributive justice perceptions in the same work groups (Li & Cropanzano, 2009; Whitman et al., 2012).

In the DT context, some employees may believe promoting DEI initiatives can influence the perceptions of the extent to which rewards they receive and how their contributions would be evaluated. It is not uncommon to observe how a majority group of employees often negatively react to policies that promote equal opportunities in the workplace. For example, researchers have found that Whites were more likely to perceive affirmative action as unfair practice because it violates the merit principle and equity principle (Bobocel et al., 1998; Kravitz, 1995). Also, other researchers demonstrated that distributive justice perceptions played a vital role in determining the levels of resistance to integrating affirmative action and willingness to accept or reject affirmative action (Leck et al., 1996). In a similar manner, Heilman and colleagues (1992) found that White men tend to judge beneficiaries of affirmative action as less competent, and less motivated with reduced career prospects. Related to DT, when employees believed DEI initiatives were deteriorating distributive justice norms, their opposition against DEI initiatives to hire more minority group members became stronger (Kidder et al., 2004). As such, employees in the organization may view DEI policies as preferential treatment, and more importantly, they may hold a similar view toward DT because DT is a vehicle to communicate core values related to DEI policies.

I argue that experiencing unfairness in distributive justice may facilitate the moral credentialing processes of trainees. That is, when trainees participate in DT with low levels of distributive justice, they will be more likely to license their previous DT, which is likely to carry moral values. Participating in DT could be conceived as ambivalent in values for some trainees as DT is, at least, a superficially moral and ethical practice, while they may still believe it as unfair practice. To counterbalance the perceptions of unfairness, some trainees will be more likely to license their previous DT experience to take advantage of the moral value DT possesses. As a result, such trainees will be more likely to express prejudice against minority group members, show a higher association with stereotypes in IAT, and show more biased behavior against DT and minority group members. I will examine how people's levels of distributive justice interact with 1) recalling a satisfying experience of DT, 2) recalling the number of previous DT, and 3) recalling the positive aspects of the current organization's DEI initiatives that will elicit stronger DT backlash via moral licensing processes. Thus, I formally hypothesize the following moderating relationships:

Hypothesis 2a: Recalling a previous experience of DT and distributive justice will interact to influence DT backlash, such that the relationship between recalling the previous experience of DT and DT backlash will be stronger when distributive justice is low.

Hypothesis 2b: Recalling the number of previous DT and distributive justice will interact to influence DT backlash, such that the relationship between recalling the number of DT and DT backlash will be stronger when distributive justice is low.

Hypothesis 2c: Recalling a positive aspect of the current organization's DEI initiatives and distributive justice will interact to influence DT backlash, such that the relationship between such recalling and DT backlash will be stronger when distributive justice is low.

In addition to distributive justice, procedural justice is another fairness perception that can negatively affect pretraining attitudes if it is violated. Procedural justice is concerned with whether the procedures are consistently, properly applied when implementing practices or policies, whether the process is free of bias, and whether the actors can have some levels of process control (Levinthal, 1980; Levinthal et al., 1980; Thibaut & Walker, 1975). Specifically, broadly defined procedural justice, which includes interpersonal and informational justice, is related to job satisfaction, OCB, trust, and job performance (see meta-analytic review: Colquitt et al., 2001). In the DT context, even if trainees believe that the core messages of DT do not violate distributive justice norms, they may still experience violations of procedural justice. For example, managers often assign employees to DT as a punishment, and trainees often learn that they were assigned to DT because their coworkers recommended it (Sanchez & Medkik, 2004). In such a case, coworkers rated trainees' behavior toward minority group members as preferential treatment. This result suggests that when DT violates procedural justice norms, it negatively impacts employees who attended training to experience more backlash but also makes other employees who did not participate in training misunderstand the trainees' behavior. More importantly, when trainees perceive the processes of DT are unfair, they will be more likely to morally license their previous DT experiences to counterbalance their feeling of unfairness. I formally posit the following moderating relationships:

Hypothesis 3a: Recalling previous experiences of DT and procedural justice will interact to influence DT backlash, such that the relationship between recalling previous experiences of DT and DT backlash will be stronger when procedural justice is low.

Hypothesis 3b: Recalling the number of previous DT and procedural justice will interact to influence DT backlash, such that the relationship between recalling the number of DT and DT backlash will be stronger when procedural justice is low.

Hypothesis 3c: Recalling a positive aspect of the current organization's DEI initiatives and procedural justice will interact to influence DT backlash, such that the relationship between such recalling and DT backlash will be stronger when procedural justice is low.

Moral Credentials and DT Backlash: The Moderating Role of Individual Differences

In this part, I posit how trainees' stable personality traits may impact the hypothesized relationships between pretraining conditions and DT backlash. Previous studies about moral licensing have focused more on the contextual factors that affect moral licensing, and the role of individual differences in the moral licensing processes has received less attention. This also has been a tendency for DT backlash. Hence, researchers and practitioners do not fully understand which types of trainees are more likely to experience DT backlash, and which trainees are more prone to general DT backlash through moral licensing processes. I introduce two individual differences moderators, social dominance orientation (SDO) and belief in a just world (BJW), in this section to test whether trainees with such personality traits are more likely to undergo DT backlash via the moral licensing process.

With regard to individual differences in moral licensing process, Liang and colleagues (2022) examined how one's propensity to morally disengage can influence moral licensing. Because moral disengagement is about deactivating the self-regulatory processes, the authors found that individuals with a high propensity to morally disengage showed pronounced moral licensing processes (Liang et al., 2022). Additionally, Zhang and colleagues (2007) found that the extent to which people are optimistic about their goals affected the licensing processes and balancing process of goal pursuit, which, in turn, increased goal-incongruent actions. With such theoretical backgrounds and empirical evidence, I contend that individual differences are likely to influence one's moral licensing processes, which are intrapersonal processes that may be affected by a person's stable ways of perceiving and construing their own moral deeds and immoral deeds. Although there is a stronger emphasis on contextual or situational predictors of moral licensing theory, some of the following studies also suggest that individual differences play a role in moral licensing processes.

Previous research regarding moral licensing often highlights the characteristics of initial moral behavior facilitating or reducing moral licensing processes. This line of research provides indirect evidence that individuals' stable personality traits can influence moral licensing process because focusing on the initial moral behavior involves a subjective interpretation of the initial behavior. For example, Conway and Peetz (2012) highlighted that recalling past moral behavior concretely elicited moral licensing compared to thinking of it abstractly. According to Conway and Peetz (2012), recalling past moral deeds concretely was related to paying attention to the act itself, whereas recalling it abstractly was related to highlighting the superordinate goals of the initial moral deeds. Relatedly, studies regarding self-regulation also found that whether individuals view their previous deeds as progress or commitment can influence moral licensing (e.g., Fishbach et al., 2006), and Mullen and Monin (2016) further interpret that such a process can influence moral licensing. Fishbach and colleagues (2009, 2014) report that if people view their previous deeds as evidence of progress, they are more likely to license their deeds and reduce efforts toward achieving a goal. On the contrary, when people were made to focus on superordinate goals and commitment toward a goal, they showed more consistent behavior toward achieving the goal. This line of evidence suggests that an individual's predisposition may influence how individuals interpret their initial moral behavior as commitment or progress, facilitating or reducing moral licensing processes. Given the direct and indirect evidence of personality traits in moral licensing processes, two individual differences moderators will provide unique perspectives on DT backlash and shed light on which types of trainees are more likely to experience DT backlash via moral licensing.

Social Dominance Orientation

In explaining the foundations of prejudice and discrimination in individuals and in society, social dominance theory (Sidanius, 1993; Sidanius & Pratto, 1999, 2004) offers an integrated perspective that centers on human nature to seek and develop a social hierarchy and intergroup conflicts produced in such processes. Unlike theories focusing specifically on psychological processes of prejudice and discrimination (e.g., motivation underlying discriminatory behavior) or social construal of the self in intergroup relations such as social identity and social categorization theories (e.g., social identity theory or social categorization theory), social dominance theory provides a multi-level perspective. The levels range from institutional foundations of prejudice and discrimination, group-level power differences between dominant and subordinate groups, to individual-level support for social hierarchy and dominance over other groups. According to the social dominance theory, forces at each level contribute to bolstering the status of dominant groups and enhancing the already existing status quo at each level. At the individual-level, dominant group members develop and become more desensitized about their prejudice and discriminatory behavior against the subordinate groups (Sidanius &

Pratto, 2004). Such development of prejudice at individual-level eventually strengthens existing prejudice and status quo at the societal level, and those become legitimizing myths (Pratto et al., 2006). Legitimizing myths, then, justify dominant groups' oppression of subordinate groups and offer privileges to dominant groups, completing the cycle of dominance. Among multiple processes and factors that promote legitimizing myths according to social dominance theory, I focus on social dominance orientation, which is individual differences that promote individual prejudice and discrimination via strong beliefs toward maintaining social hierarchy (Pratto et al., 1994), because it can significantly influence moral licensing processes of DT backlash and shed light on who has a greater likelihood to experience such processes along with unfairness perceptions toward DT.

According to Pratto and colleagues (1994), social dominance orientation (SDO) is defined as an individual preference for superiority and dominance of the group to which a person belongs. As individual-level personality trait that is related to the development of legitimizing myths, SDO highlights the extent to which a person prefers social stratification based on power and status over equal status. As a result, individuals with a high level of SDO support ideologies and myths that strengthen group inequality and tend to view possessing prejudice as natural and conveying truth about group characteristics (Pratto et al., 1994). As a result, they tend to score high on prejudice against various minority categories, such as Blacks (Pratto et al., 1994; Sidanius & Pratto, 2001), and women (Sidanius et al., 1996). Moreover, because individuals with high SDO are inclined to assume that it is legitimate for the superior group members to dominate and oppress subordinate group members, SDO is highly correlated with sexism, and racism (Sidanius & Pratto, 1993). On the other hand, individuals with low SDO value group equality and resist group differentiation based on power or status differences. Taken together, individuals with high SDO tend to support legitimizing myths and believe that having and demonstrating prejudice is socially acceptable for dominant groups.

Sidanius and Pratto (2004) theorized that there are four conditions that can increase one's level of SDO. Those are group membership, socialization experiences, temperamental personality traits, and gender. According to these theorists, whether an individual belongs to a dominant group or subordinate group is a critical condition that determines one's level of SDO, and the dominant group membership tends to develop SDO. Individual socialization experiences in different life domains such as education, occupation, and religion can enhance or inhibit the level of SDO. One's stable personality traits, such as narcissism and empathy, can also influence the development of SDO. Finally, being a male is related to possessing significantly higher levels of SDO than women.

I maintain that trainees' levels of SDO may moderate the hypothesized relationships between previous experiences of DT and DT backlash. The hypothesized relationship that highlights moral licensing processes in DT backlash could vary by trainees' levels of SDO because trainees with high SDO may readily and strategically use the positive or moral aspects of DT to morally license themselves compared to trainees with low SDO. I contend that morally licensing their previous experiences of DT could become a viable route to express their existing prejudice and beliefs about social inequality. This is possible because high SDO individuals perceive diversity policies such as affirmative action as a threat to their dominant group status (Sidanius et al., 1996), and researchers have found that a threat to a group that individuals belong to can increase private prejudice against the minority group via moral licensing (Effron & Knowles, 2015). This tendency was strengthened as the participants believed that their groups were entitative (coherent and unified wholes). Individuals with high SDO may view dominant groups as more unified and, therefore, perceive DT as a threat to their groups. Taken together, individuals with high SDO will be more likely to strategically use their previous DT experiences as moral credentials leading to DT backlash, and this tendency will be exacerbated when they also hold unfair perceptions about DT. As such, I posit the following hypotheses:

> *Hypothesis 4a: Recalling previous experiences of DT and SDO will interact to influence DT backlash, such that the relationship between recalling previous experiences of DT and DT backlash will be stronger when SDO is high.*
>
> *Hypothesis 4b: Recalling the number of previous DT and SDO will interact to influence DT backlash, such that the relationship between recalling the number of DT and DT backlash will be stronger when SDO is high.*
>
> *Hypothesis 4c: Recalling a positive aspect of the current organization's DEI initiatives and SDO will interact to influence DT backlash, such that the relationship between such a recalling and DT backlash will be stronger when SDO is high.*

Belief in a just world (BJW)

BJW is a personal tendency to believe that the social world functions in a fair and just way because people are getting what they deserve (Furnham & Procter, 1989; Lerner, 1965, 1980). On the one hand, BJW can provide individuals with assumptions that the world that their daily lives occur is stable and orderly to some extent, which, in turn, eliminates some levels of uncertainty and makes them focused on achieving long-term goals (Lerner & Miller, 1978). On the other hand, because BJW also offers assumptions that the world is already just, people who hold strong beliefs about it tend to blame or derogate victims and believe that some people do not receive rewards because they do not make enough effort (Feather, 1984; Furnham, 2003; Lerner, 1980). Individuals with high BJW are more likely to believe that the current social or political system is fair and correct. As a result, researchers found that BJW is correlated with political conservatism, authoritarianism, Protestant work ethics, and internal locus of control (Furnham & Procter, 1989). Based on the rigid beliefs about the overall system, individuals with high levels of BJW tend to show higher levels of overall prejudice (Staub, 1996), prejudice against Blacks (Rim, 1988), and against people with depression (Crandall & Cohen, 1994). Relatedly, people with a strong BJW are less likely to demonstrate sympathy toward poverty, poor people (Furnham & Gunter, 1984; Wagstaff, 1984), and feminism (Rubin & Peplau, 1975). Consequently, people with a strong belief in a just world are inclined to attribute victims of financial failure to personal incompetence rather than accusing the systemic causes in society or environmental constraints. Taken together, BJW is one of the individual differences that may increase DT backlash as individuals with this belief may justify racial inequality and discrimination.

In the DT context, individuals who hold a strong belief in a just world may have negative attitudes against implementing DT in organizations because, for them, DT unnecessarily promotes changes in an already just workplace. They might further believe that receiving DT once or twice is sufficient for managing diversity in the workplace and that their organization is already doing enough to maintain a fair workplace (Dobbin & Kalev, 2018). As a result, when those people are introduced to DT, they might recall only a few numbers of DT to justify that their beliefs about their organizations are reasonable. Based on the overly simplified belief that the world is a fair place, they might overly acknowledge that their organization is already treating employees without any prejudice. This provides more opportunities for such people to license their own experiences of DT and their organization's overall DEI initiatives. When this occurs, employees may believe that their organization is ethical and pay less attention to monitoring discriminatory behavior in the organization and regulate their behavior (Leslie, 2019). Trainees with high BJW may engage in DT at a superficial level because they are likely to believe DT and DEI initiatives are unnecessary. Such superficial participation in DT may intensify the moral licensing processes because it will highlight the progress of their good deeds rather than commitment or superordinate goals of moral deeds (i.e., DT). Results from previous research also support that when the initial moral behavior highlights progress rather than commitment, individuals are more likely to go through moral licensing (Fishbach et al., 2006, 2009; Zhang et al., 2007). Taken together, the hypothesized relationship between pretraining conditions and moral credentialing processes will be moderated by the levels of trainees' BJW. Based on these rationales, I formally posit the following three-way interaction hypotheses.

Hypothesis 5a: Recalling previous experiences of DT and BJW will interact to influence DT backlash, such that the relationship between recalling previous experiences of DT and DT backlash will be stronger when BJW is high.

Hypothesis 5b: Recalling the number of previous DT and BJW will interact to influence DT backlash, such that the relationship between recalling the number of DT and DT backlash will be stronger when BJW is high.

Hypothesis 5c: Recalling a positive aspect of the current organization's DEI initiatives and BJW will interact to influence DT backlash, such that the relationship between such recalling and DT backlash will be stronger when BJW is high.

STUDY 1

In Study 1, I examined whether people who recall previous experiences of DT report higher levels of DT backlash compared to people who did not recall such experiences. Because Study 1 was exploratory in nature, I only examined whether recalling previous experiences influenced participants' emotions, attitudes, and behavioral tendencies without administering any type of DT. I also examined how participants' justice perceptions and individual differences (i.e., SDO, BJW) moderated the direct effects of recalling previous DT experiences on DT backlash. For dependent variables, which were the various manifestations of DT backlash, I measured moral credentials, moral credits, psychological entitlement, discrete emotions, implicit bias, explicit prejudice, willingness to hire a minority group member, and willingness to help minority group members. Table 3 in Appendix C describes the experimental design for Study 1.

Method

Participants

The sample size was determined before data collection began at 400. The sample size was decided based on the power analysis (G*Power analysis). The overall effect size of moral licensing provided by the meta-analysis, (Blanken et al., 2015) was 0.31 (Cohen's *d*). G*Power analysis suggests collecting a total of 400 subjects to acquire 80% statistical power (Cohen, 1992; $\alpha = 0.5$, $d = 0.31$).

I collected samples from Prolific, which is an online survey platform. To only include participants with previous DT experiences, I administered a screening procedure that asked questions about whether they had participated in DT previously, whether they participated in DT offered by their current employers, and whether they are currently hired as full-time employees. Participants who were not full-time employees residing in North America or did not have any

DT experiences were not invited to the main study. A total of 1,101 participants were recruited to participate in the screening procedure for Study 1, and 710 participants passed the screening process. Among those, 455 participants joined the main study, and 394 participants successfully completed the main part of Study 1. Fourteen participants who failed to pass the attention checks or manipulation checks were excluded from the analyses. The final number of participants included in the analysis for Study 1 was 380. Each participant received $0.40 for the screening procedure and $5.50 for completing the main experiment. Participants were randomly assigned to one of the four conditions.

Procedure

Upon passing the screening procedure, participants received invitations to join the main study. After agreeing to the informed consent to participate in the experiment, participants were randomly assigned to four different groups, which are comprised of three manipulation groups, and one control group. The three manipulation groups were asked to recall either 1) their previous experiences of taking DT, 2) their previous number of DT, or 3) the positive aspects of their organization's current DEI initiatives. Participants in each condition were asked to answer multiple open-ended questions, and those questions were related to specific aspects of DT and how DT impacted their behavior in the workplace. Participants in the first manipulation condition were asked to answer how much they liked previous DT, what they were able to learn after DT, and how did DT influence their personal and professional relationships with colleagues from different backgrounds. Participants in the second manipulation condition were requested to provide the number of DT they participated in college, at their previous workplace, at their current workplace, and the number of DT they have taken in total. Participants in the third manipulation condition first provided the ongoing DEI policies at their current workplace, and then, shared how such policies provided a chance to better understand colleagues from different backgrounds, improve equity in the workplace, and interact better with other colleagues. Participants in a control condition were not asked to recall anything. Participants in every condition first performed implicit association tests (IAT) and went through justice perceptions questionnaires previous to manipulation. After the manipulation, participants performed the second round of IATs and answered survey questionnaires related to DT backlash manifestations. Participants were fully debriefed about why they were asked to recall previous DT experiences and were given survey completion codes to enter to receive monetary rewards.

Measures

Participants rated the extent to which they agreed with the statements on a 5-point Likert scale otherwise specified. Every scale used in this study can be found in Appendix E.

Moral credentials. To measure moral credentials, I followed Loi et al. (2020) and adapted the five-item internalization scale of moral self-regard and the extent to which participants perceive their moral identity. This scale was originally developed by Aquino and Reed (2002), and I adapted it to match the DT context. Sample items are "It would make me feel good to be a person who has these characteristics," and "It would make me feel good to be a person who has these characteristics." Cronbach's alpha for this scale was .91.

Moral credits. To measure moral credits for DT backlash, I have created a five-item scale, following a scale developed by Lin and colleagues (2016). To address the DT context, I considered that adapting the existing scale was not sufficient and developed the five items. Sample items are "I feel like I have earned recognition for becoming a good person by participating in diversity training," "My participation in diversity training is evidence that I am a good person." Cronbach's alpha for this scale was .95.

Psychological entitlement. I also measured psychological entitlement as one of the state measures that are often elicited by previous good deeds (e.g., Loi et al., 2020). I used the nine-item scale developed by Campbell and colleagues (2004). Sample items are “I honestly feel I’m just more deserving than others,” “I demand the best because I’m worth it.” Cronbach’s alpha for this scale was .84.

DT backlash attitudes – explicit prejudice against minority group members. I measured the explicit prejudice against minority groups using the Modern Racism Scale (McConahay, 1986), which asks subtle questions about one’s attitudes toward African Americans. Cronbach’s alpha for this scale was .75.

DT backlash affects – discrete emotions. To measure discrete emotions, I draw on Chung et al. (2022)’s short form of emotional scale called facets of emotional experiences in everyday life scale, which uses 12 items. It measures six facets, and those are anger, fear, joy, love, sadness, and surprise. Two items belong to each facet. For anger, I asked the participants to rate how they felt after recalling the previous experience of DT, using 5-point Likert scale that was anchored not at all = 1 to extremely = 5. Cronbach’s alpha for the discrete emotions are as follows: anger = .94, fear = .86, joy = .95, love = .97, sadness = .88, and surprise = .95.

DT backlash attitudes – implicit association tests (IAT). The IAT is the test that measures response latencies of implicit categories and assesses the strengths of automatic associations between categories by going through computer-based categorization tasks (Greenwald et al., 2009). For example, when an individual is quicker to associate White and positive features than associating African American and positive features on a series of categorization tasks, then the test shows one’s preference for Whites to Blacks. Despite some critiques toward using the IAT due to questionable reliability and validity (Blanton et al., 2009), Greenwald and colleagues

(2009) have provided meta-analytic findings that demonstrate incremental and predictive validity of implicit and explicit prejudice on behavior. See Appendix E – for photos and words used for IAT. Cronbach's alpha for pretest IAT was .87 and for posttest IAT was .84.

DT backlash behavior – willingness to hire a White person for a job over a Black person. Following Monin and Miller (2001), I measured trainees' willingness to hire a White person over a Black person using a single-item scale based on the scenario that describes whether selecting a person with a particular ethnic background is more appropriate.

DT backlash behavior – willingness to help minority group members. I adapted and used Williams and Anderson's (1991) seven-item OCB-I scale to measure the extend to which participants are willing to help minority group members. Sample items are "I would help racial minority group members who have heavy workloads" and "I would pass along information to racial minority group members." Cronbach's alpha for this scale was .90.

Justice Perceptions – distributive justice. I adapted four items from the distributive justice scale that was developed and validated by Colquitt et al. (2001). Distributive justice is concerned with how people view the outcomes based on their contribution. People may readily perceive DEI initiatives rather than DT as something that negatively impacts the ratio of outcomes to contribution. Therefore, I adapted the distributive justice scale to reflect DEI initiatives influencing trainees' perceptions of unfairness about rewards and outcomes. Sample items for distributive justice are "Promoting DEI initiatives at your organization helps secure outcomes that reflect the effort you have put into your work?", "Promoting DEI initiatives at your organization is fair for your work outcomes given your performance?" Cronbach's alpha for this scale was .94.

Justice Perceptions – Procedural justice. To measure procedural justice, I adapted the seven-item scale developed and validated by Colquitt et al. (2001). I measured two different facets of procedural justice perceptions: the extent to which trainees feel about processes during DT, and the extent to which trainees feel about processes of assigning employees to DT. Sample items for procedural justice for the former are "Have you been able to express your views and feelings when your diversity training was assigned to you?", "Have those procedures for assigning you to diversity training been applied consistently?" Samples items for the latter are "Have you been able to express your views and feelings during diversity training?", "Have those procedures during diversity training been applied consistently?" Cronbach's alpha for the assignment was .84, and Cronbach's alpha for procedural justice during DT was .85.

Individual differences – Social dominance orientation. To measure one's SDO, I used an 8-item scale developed and validated by Ho et al. (2015). In this measurement, the researchers introduced two subdimensions of SDO – SDO-Dominance (SDO-D), which refers to "a preference for group-based dominance hierarchies in which dominant groups actively oppress subordinate groups," and SDO-Egalitarian (SDO-E), which represents one's "opposition to equality between groups, as supported by an interrelated network of subtle hierarchy-enhancing beliefs and social policies" (Ho et al., 2015, pp. 1004). Cronbach's alpha for this scale was .94.

Individual differences – Belief in a just world. I measured trainees' BJW using the 5-item scale created by Lipkus (1991), which is the 20-item version that improved from the limitations of the formerly popular scale developed by Rubin and Peplaus (1975). However, I followed Judge et al.'s instructions (1998) to utilize a shorter version of BJW. Sample items are "Basically, the world is a just place," "By and large, people deserve what they get." Cronbach's alpha for this scale was .58.

Control variables. I controlled for trainees' trait psychological reactance to eliminate alternative explanations that may explain DT backlash. *Trait psychological reactance* was measured with a twelve-item scale (e.g., "I become angry when my freedom of choice is restricted," "It irritates me when someone points out things which are obvious to me") developed by Hong and Page (1989) and validated by Hong and Faedda (1996). Cronbach's alpha for procedural justice during DT was .85.

Analysis

I conducted ANOVA and ANCOVA to compare the group mean differences of DT backlash between the treatment groups and the control group. To perform moderator analyses, I dummy-coded each manipulation group, and the referent group for the regression analysis was the control group, which did not recall anything. I used multiple regression and hierarchical regression because the moderator variables used in this study were continuous variables and I test the interaction between categorical and continuous variables. Because the results of moderator analyses were based on the dummy variables, unstandardized regression coefficients represent the mean differences from those of the referent group in terms of DVs, and intercepts represent the means of the referent groups' DVs (Cohen et al., 2013). I mean-centered all the moderators, and because the independent variables were categorical, I did not center them.

Results

Table 6 in Appendix F shows descriptive statistics and correlation coefficients for Study 1, respectively. Fourteen participants who failed to pass the attention check were excluded from the analyses.

Main Effects of Manipulations on DT Backlash

Figure 1 through 3 in Appendix G describes the results of ANOVA. The results of one-way ANOVA showed that the effects of manipulations were not statistically significant on DT backlash variables, including psychological entitlement, implicit prejudice, explicit prejudice, and willingness to help or hire minority group members, thereby not supporting Hypotheses 1a, 1b, and 1c. Among six discrete emotions (i.e., anger, fear, joy, love, sadness, and surprise), the group differences only existed for anger $F(1, 388) = 2.54$, $p < .10$, $\eta^2 = .02$ (manipulation group 1 $M = 1.68$ $SD = 1.05$, manipulation group 2 $M = 1.55$ $SD = 0.85$, manipulation group 3 $M = 1.39$ $SD = 0.75$, control group $M = 1.72$, $SD = 1.72$). Tukey's HSD post hoc analysis showed that the mean difference between the group that recalled the positiveness of DEI policies at the current workplace and the control group was (*M* difference = -.34, SE = 0.14) at $p < .10$.

Also, the effects of manipulations on moral credentials and moral credits were significant at $p < .10$ level. For moral credentials, means between groups were significantly different $F(1, 388) = 2.20$, $\eta^2 = .02$ at $p < .10$ level (manipulation group 1 $M = 3.65$, $SD = 1.12$; manipulation group 2 $M = 3.83$, $SD = 1.00$; manipulation group 3 $M = 3.90$, $SD = 0.85$; control group $M = 3.56$, $SD = 1.15$). Tukey's HSD test revealed that the mean difference between the group that recalled the positiveness of DEI policies at the current workplace and the control group was at $p = .110$.

There were also group mean differences regarding moral credits $F(1, 388) = 2.13$, $p = .10$, $\eta^2 = .02$. Tukey's HSD post hoc analysis showed that participants in manipulation group 3 ($M = 3.01$, $SD = 0.99$) who recalled positive aspects of DEI policies at their current workplace showed higher levels of moral credits than manipulation group 1 ($M = 2.63$, $SD = 1.19$; *M* diff =

0.39, *SE* = .16, *p* < .10; manipulation group 1 *M* = 2.62 *SD* = 1.19, manipulation group 2 *M* = 2.86 *SD* = 1.02, manipulation group 3 *M* = 3.02 *SD* = 0.99, control group *M* = 2.84, *SD* = 1.11)

I also ran regression analyses to test the significance of mean differences between each group using the *t* test. Results of the regression analyses on moral credentials demonstrated that the mean differences between manipulation group 2 (*b* = .27, *SE* = .15, *t* = 1.83, *p* < .10) and the control group (*intercept* = 3.56, *SE* = .11, *t* = 33.91, *p* < .001), and between manipulation group 3 (*b* = .34, *SE* = .15, *t* = 2.26, *p* < .05) and the control group were significant. For anger, the mean difference between manipulation group 3 (*b* = -.34, *SE* = .14, *t* = 2.50, *p* < .05) and the control group (*intercept* = 1.72, *SE* = .09, *t* = 18.28, *p* < .001) was significant. Regarding explicit prejudice, the mean of manipulation group 2, who recalled the number of previous DT (*b* = -.21, *SE* = .09, *t* = 2.33, *p* < .05), was significantly different from the mean of the control group (*b* = 2.16, *SE* = .07, *t* = 30.86, *p* < .001). Also, recalling the positive aspects of DEI policies tended to have lower means for explicit prejudice (*b* = -.16, *SE* = .10, *t* = 1.60, *p* < .10) and hiring decisions for minority group members (*b* = .37, *SE* = .20, *t* = 1.99, *p* < .05) compared to those of the control group (explicit prejudice: *b* = 2.16, *SE* = .07, *t* = 30.86, *p* < .001; hiring decisions: *b* = 3.96, *SE* = .13, *t* = 30., *p* < .001). Table 7 in Appendix G summarizes the regression analysis results.

Moderating Effects of Distributive Justice Perceptions

Table 8 in Appendix H and Figures 4 to 9 in Appendix I summarize the moderation analysis results of distributive justice. Hypothesis 2 was concerned with the moderating effect of distributive justice perceptions on the relationships between manipulations and DT backlash, and I hypothesized that possessing lower levels of distributive justice would strengthen the relationship. Distributive justice perceptions had direct effects on anger (*b* = -.28, *SE* = .04, *t* =-

7.15, *p* < .001), moral credentials (*b* = .79, *SE* = .07, *t* =11.04, *p* < .001), moral credits (*b* = .48, *SE* = .09, *t* = 5.33, *p* < .001), explicit prejudice (*b* = -.32, *SE* = .03, *t* = -10.67, *p* < .001), and OCB (*b* = .23, *SE* = .03, *t* = 3.67, *p* < .01). Interactions terms between manipulations and distributive justice were significant on anger, moral credentials, and explicit prejudice. For anger, the product term between the manipulation group 1 and distributive justice was significant (*b* = .20, *SE* = .11, *t* =1.82) at *p* < .10 level. Specifically, the negative relation of manipulations on anger was stronger for individuals who are lower (vs. higher) in distributive justice. A simple slope analysis revealed that both slopes were not statistically significant (low distributive justice *b* = -.54, *t* = -1.29, *ns*; high distributive justice *b* = -.09, *t* = -0.21, *ns*). For moral credentials, product terms of manipulation groups and distributive justice were statistically significant for all three manipulation groups (manipulation group 1 *b* = -.20, *SE* = .09, *t* = -2.15, *p* < .05; manipulation group 2 *b* = -.23, *SE* = .10, *t* = -2.31, *p* < .05; manipulation group 3 *b* = -.31, *SE* = .11, *t* = -2.93, *p* < .01). Specifically, the positive relation between manipulation and moral credentials was stronger for individuals who are lower (vs. higher) in distributive justice. A simple slope analysis showed that manipulation 1 had a positive relation with moral credentials when distributive justice was low (*b* = .48, *SE* = .15, *t* = 3.26, *p* < .01), but the relation was not significant when distributive justice was high (*b* = .02, *t* = 0.11, *ns*); manipulation 2 had stronger relation with moral credentials when distributive justice was low (*b* = .42, *SE* = .17, *t* = 2.51, *p* < .05) but not significant when distributive justice was high (*b* = -.11, *t* = -0.72, *ns*); manipulation 3 had stronger relation with moral credentials when distributive justice was low (*b* = .67, *SE* = .17, *t* = 4.03, *p* < .001) but not significant when distributive justice was high (*b* = -.05, *t* = -0.29, *ns*).

For explicit prejudice, the product term of manipulation groups between distributive justice was only significant for manipulation group 2 (*b* = .22, *SE* = .08, *t* = 2.80, *p* < .01). A product term of manipulation group 3 and distributive justice was significant at *p* < .10 level (*b* = .14, *SE* = .08, *t* = 1.73, *p* < .10). A simple slope analysis revealed that manipulation 2 had a negative relation with explicit prejudice when participants had lower levels of distributive justice *b* = -.74, *SE* = .23, *t* = -3.29, *p* < .01), but this relation was not significant when distributive justice was high (*b* = .14, *t* = 1.02, *ns*). The simple slopes from manipulation 3 and distributive justice perceptions were not statistically significant (low distributive justice *b* = -.31, *t* = -1.02, *ns*; high distributive justice *b* = .02, *t* = 0.13, *ns*).

Overall, contrary to my hypothesis 2, manipulations reduced DT backlash, and such effects of my manipulations on DT backlash were stronger for participants with low distributive justice perceptions.

Moderating Effects of Procedural Justice Perceptions

Tables 9 and 10 in Appendix H and Figures 10 to 14 in Appendix I summarize the moderation analysis results of procedural justice. I measured two facets of procedural justice. The extent to which trainees believe that their assignment to DT was fair in terms of procedure had direct effects on anger (*b* = -.33, *SE* = .05, *t* =-6.56, *p* < .001), moral credentials (*b* = .31, *SE* = .04, *t* = 8.90, *p* < .001), moral credits (*b* = .50, *SE* = .06, *t* = 8.91, *p* < .001), explicit prejudice (*b* = -.17, *SE* = .04, *t* = -4.60, *p* < .001), OCB (*b* = .29, *SE* = .04, *t* = 8.45, *p* < .001), and willingness to hire minority group members (*b* = -.20, *SE* = .07, *t* = -2.74, *p* < .01). Hypothesis 3 was concerned with the moderating effects of procedural justice perceptions on the relationships between manipulations and DT backlash. The results showed that the moderating effects of procedural justice of the assignment of DT were only statistically significant for moral credentials. For moral credentials, the product term of manipulation group 3 and procedural justice was statistically significant (*b* = -.28, *SE* = .11, *t* = -2.52, *p* < .05). A simple slope analysis demonstrated that manipulation 3 had a positive relation with moral credentials when procedural justice was low (*b* = .65, *SE* = .24, *t* = 2.69, *p* < .01) but not significant when procedural justice was high (*b* = .07, *t* = 0.52, *ns*). This result showed that Hypothesis 3 was not supported, as the interaction between manipulation 3 and procedural justice reduced DT backlash.

Another facet of procedural justice, which was about justice perceptions regarding procedure during DT, had direct effects on anger (*b* = -.34, *SE* = .05, *t* =-2.81, *p* < .01), moral credentials (*b* = .31, *SE* = .03, *t* = 9.13, *p* < .001), moral credits (*b* = .51, *SE* = .06, *t* = 9.15, *p* < .001), explicit prejudice (*b* = -.22, *SE* = .04, *t* = -6.14, *p* < .001), OCB (*b* = .30, *SE* = .03, *t* = 9.02, *p* < .001), and willingness to hire minority group members (*b* = -.18, *SE* = .07, *t* = -2.41, *p* < .05). The moderating effects of procedural justice during DT were statistically significant for moral credentials, explicit prejudice, and OCB. For moral credentials, the product term of manipulation group 3 and procedural justice was statistically significant (*b* = -.32, *SE* = .11, *t* = -3.01, *p* < .05). A simple slope analysis revealed that manipulation 3 had a positive relation with moral credentials when participant's procedural justice was low (*b* = .79, *SE* = .19, *t* = 4.11, *p* < .001), while not significant for high procedural justice (*b* = -.03, *t* = -0.14, *ns*). For explicit prejudice, the product term of manipulation group 2 and procedural justice was statistically significant (*b* = .20, *SE* = .10, *t* = 1.99, *p* < .05). The product term of manipulation group 3 and procedural justice on explicit prejudice (*b* = .19, *SE* = .11, *t* = 1.74) was statistically significant at *p* < .10 level. A simple slope analysis showed that manipulation 2 had a negative relation with explicit prejudice when participant's procedural justice was low (*b* = -.41, *SE* = .13, *t* = -3.22, *p* < .01), while not significant for high procedural justice (*b* = -.05, *t* = -0.39, *ns*); manipulation 3 had a negative relation with explicit prejudice when participant's procedural justice was low (*b* = -.35, *SE* = .14, *t* = -2.58, *p* < .05), while not significant for high procedural justice (*b* = -.01, *t* = -0.02, *ns*).

For OCB, the product term of manipulation group 3 and procedural justice was statistically significant on OCB (*b* = -.23, *SE* = .11, *t* = -2.19, *p* < .05). A simple slope analysis demonstrated that manipulation 3 had a negative relation with OCB when participant's procedural justice was high (*b* = -.25, *SE* = .13, *t* = -1.92) at *p* < .10 level, while not significant for low procedural justice (*b* = .17, *t* = 1.29, *ns*). These results show that the interactions between manipulations and procedural justice perceptions lessened DT backlash while strengthening DT backlash regarding OCB. Thus, Hypothesis 3 was partially supported.

Moderating Effects of Personality Traits

I also conducted regression analyses to test the moderating effects of SDO and BJW between the manipulations and DT backlash. Table 11 in Appendix H and Figures 15 to 17 in Appendix I summarize the moderation analysis results of SDO. SDO had direct effects on anger (*b* = .42, *SE* = .13, *t* =-2.31, *p* < .001), moral credentials (*b* = -.46, *SE* = .04, *t* = -10.80, *p* < .001), moral credits (*b* = -.16, *SE* = .08, *t* = -2.08, *p* < .05), psychological entitlement (*b* = .22, *SE* = .05, *t* = 4.44, *p* < .001), explicit prejudice (*b* = .65, *SE* = .07, *t* = 18.75, *p* < .001), and OCB (*b* = -.41, *SE* = .04, *t* = -9.51, *p* < .001). SDO also directly affected willingness to hire minority group members (*b* = .18, *SE* = .10, *t* = 1.88) at *p* < .10 level. Hypothesis 4 was concerned with the moderating effect of SDO on the relation between manipulations and DT backlash. The moderating effects of SDO were not statistically significant on DT backlash except for implicit prejudice and psychological entitlement. For implicit prejudice, recalling the number of DT taken interacted with SDO (*b* = -2.05, *SE* = 1.10, *t* = 2.01, *p* < .05) to reduce implicit bias compared to the control group. A simple slope analysis showed that manipulation 2 had a negative relation with implicit prejudice when participant's SDO was high (*b* = -2.53, *SE* = 1.02, *t* = -2.49, *p* < .05), while not significant for low procedural justice (*b* = .32, *t* = 0.31, *ns*).

For psychological entitlement, the product term of manipulation group 2 and SDO on psychological entitlement (*b* = .28, *SE* = .14, *t* = 2.09, *p* < .05) was statistically significant. Also, the product term of manipulation group 3 and SDO on psychological entitlement (*b* = .26, *SE* = .14, *t* = 1.85) was statistically significant at *p* < .10 level. A simple slope analysis showed that both slopes for SDO were not significant for manipulation 2 (low SDO *b* = -.18, *t* = -1.35, *ns*; high SDO *b* = .21, *t* = 1.58, *ns*). However, for manipulation 3, the result showed that it had a positive relation with psychological entitlement when participant's SDO was high (*b* = .32, *SE* = .13, *t* = 2.45, *p* < .05), while not significant for low SDO (*b* = -.04, *t* = -0.30, *ns*). Because the interactions between manipulations and SDO increased psychological entitlement while reducing implicit prejudice, Hypothesis 4 was partially supported.

Table 12 in Appendix H and Figures 18 to 22 in Appendix I summarize the moderation analysis results of BJW. BJW had direct effects on moral credentials (*b* = -.18, *SE* = .05, *t* = -3.67, *p* < .001), moral credits (*b* = -.23, *SE* = .08, *t* = -2.73, *p* < .01), explicit prejudice (*b* = .09, *SE* = .05, *t* = 1.80, *p* < .10), and OCB (*b* = -.17, *SE* 1= .05, *t* = -3.35, *p* < .01). Hypothesis 5 was concerned with the moderating effect of BJW on the relation between manipulations and DT backlash. The moderating effects of BJW were statistically significant for moral credentials, psychological entitlement, and explicit prejudice. Regarding moral credentials, the product term of manipulation group 2 (*b* = .24, *SE* = .15, *t* = 1.67) and BJW was significant at *p* < .10 level. A simple slope analysis showed that manipulation 2 had a positive relation with moral credentials when participant’s BJW was high (*b* = .62, *SE* = .21, *t* = 2.98, *p* < .01), while not significant for low BJW (*b* = -.05, *t* = -0.24, *ns*).

With respect to psychological entitlement, the product term of manipulation group 3 (*b* = -.25, *SE* = .16, *t* = -1.90) was significant at *p* < .10 level. The simple slopes from manipulation 3 and BJW on psychological entitlement were not statistically significant (low BJW *b* = .19, *t* = 1.37, *ns*; high BJW *b* = -.20, *t* = -1.36, *ns*). For explicit prejudice, the product terms of manipulation group 2 (*b* = -.30, *SE* = .15, *t* = -2.04, *p* < .05), and manipulation group 3 (*b* = -.33, *SE* = .16, *t* = -2.13, *p* < .05) were statistically significant. The product term of manipulation group 1 and BJW (*b* = -.27, *SE* = .15, *t* = -1.85) was significant at *p* < .10 level. The simple slopes from manipulation 3 and distributive justice perceptions were not statistically significant (low distributive justice *b* = -.31, *t* = -1.02, *ns*; high distributive justice *b* = .02, *t* = 0.13, *ns*). Again, a simple slope analysis showed that manipulation 1 had a negative relation with explicit prejudice when participant’s BJW was high (*b* = -.31, *SE* = .13, *t* = -2.32, *p* < .05), while not significant for low BJW (*b* = .05, *t* = 0.34, *ns*); manipulation 2 had a negative relation with explicit prejudice when participant’s BJW was high (*b* = -.42, *SE* = .13, *t* = -3.13, *p* < .01), while not significant for low BJW (*b* = -.03, *t* = -.19, *ns*); manipulation 3 had a negative relation with explicit prejudice when participant’s BJW was high (*b* = -.38, *SE* = .14, *t* = -2.73, *p* < .01), while not significant for low BJW (*b* = .05, *t* = 0.39, *ns*). Contrary to my prediction, the interaction between manipulations and BJW lessened the DT backlash. As a result, it failed to support Hypothesis 5.

Supplemental Analyses

To examine whether the present findings withstand more robust testing, I entered psychological reactance, which is one of the strongest psychological mechanisms of DT backlash, as a control variable. All of the results remain unchanged except for the moderation effects of manipulation group 3 and SDO on psychological entitlement.

Discussion

Table 19 in Appendix N summarizes all the significant results of Study 1. Contrary to my predictions, the statistical results showed that the manipulations (i.e., recalling previous experiences of DT, recalling the number of DT taken, or the positiveness of the DEI policies at the current workplace) were significant on anger, moral credentials, and moral credits. Rather than generating moral licensing effects, the manipulations created consistency effects on those DT manifestations. That is, the manipulations reduced anger and moral credits while increasing participants' moral credentials compared to those of the control group. The results of ANOVA show that manipulations created mean differences between groups on moral credits. There was a mean difference between manipulation groups 1 and 3, and it showed that participants who recalled the positiveness of the DEI policies reported higher levels of moral credits than participants who recalled the previous experiences of DT. These results indicate that such manipulations increased some levels of moral licensing processes for participants in manipulation group 3 compared to people in manipulation group 1.

The regression analyses based on t-tests, which compared the manipulation group's means to the means of the control group, further revealed that the manipulations generated consistency effects. Specifically, recalling the number of DT or the positiveness of the DEI policies at the current workplace reduced anger and explicit prejudice while increasing moral credentials. Although there were statistical results that indicate the manipulations created moral licensing effects, consistency effects were observed in various DT backlash manifestations. Contrary to my expectation, recalling the number of previous DT taken or the positiveness of the

DEI policies reduced explicit prejudice compared to the control group. Also, recalling the positiveness of DEI reduced participants' anger compared to the control group. These results indicate that the manipulations were very powerful in terms of lessening the key manifestations of DT backlash.

On the other hand, participants who recalled the positiveness of the DEI policies at the current workplace increased the likelihood of hiring majority group members, showing that such manipulation can also produce moral licensing effects. This result suggests that the manipulations can differently affect manifestations of DT backlash as recalling the positiveness of the DEI policies lessened the attitudinal manifestations (i.e., explicit prejudice), whereas it increased the behavioral manifestations of DT backlash (i.e., willingness to hire minority group members).

However, the moderation analysis further revealed that the manipulations effectively lessened DT backlash manifestations. The interaction between manipulations and distributive justice perceptions showed that manipulations were particularly effective for participants who perceive DEI initiatives are unfair in preventing DT backlash manifestations. Specifically, participants who showed lower levels of distributive justice and underwent manipulations were more likely to gain moral credentials and reduce anger and explicit prejudice compared to who did not go through manipulations. When participants with lower levels (vs. higher) of distributive justice recalled the past experiences of DT, they reduced anger and gained more moral credentials. When participants with lower levels (vs. higher) of distributive justice either recalled the number of DT attended or the positive aspects of DEI policies, they were more likely to obtain moral credentials and reduce explicit prejudice. Compared to participants with lower distributive justice perceptions, participants with higher justice perceptions also showed that they had higher levels of moral credentials and lower levels of anger and explicit prejudice regardless of experiencing manipulations. This tendency indicates that the manipulations did not trigger moral licensing effects for participants who perceive DEI values are fair in terms of distribution of outcomes, but the manipulations can lessen DT backlash of participants who perceive DEI values are unfair.

The moderating role of procedural justice perceptions revealed similar patterns as participants possessing lower procedural justice perceptions about the assignments of DT increased moral credentials when they recalled previous experiences of DT. Moreover, when participants with lower procedural justice perceptions regarding procedure during DT recalled either the number of DT taken or the positiveness of DEI policies, they were able to report lower levels of explicit prejudice. In the same manner as distributive justice perceptions, participants who had higher levels of procedural justice perceptions showed higher levels of moral credentials and lower levels of explicit prejudice regardless of the group conditions. Thus, this finding also strongly suggests that the manipulations lessened DT backlash, particularly for participants with unfairness perceptions.

Furthermore, the moderation analyses regarding personality also showed that the manipulations effectively reduced DT backlash manifestations for participants who are more likely to be prejudiced against minority groups. When participants with high BJW experienced manipulations, they reduced explicit prejudice. Individuals with lower BJW showed a lower baseline prejudice, and they maintained low levels of prejudice irrespective of manipulations. Additionally, when participants with high BJW recalled the number of DT, they reported higher moral credentials than those who did not go through manipulation. Thus, recalling experiences of

DT or the positiveness of DEI policies also functioned as a buffer against DT backlash manifestations for those who possessed stable personality traits.

On the other hand, the moderation analysis of SDO revealed an interesting finding. The analysis showed that, unlike BJW or justice perceptions, the manipulations did not interact with SDO to reduce DT backlash manifestations such as anger or explicit prejudice. However, when participants with high SDO recalled the number of DT they attended, it reduced their implicit prejudice. At the same time, those participants increased psychological entitlement after recalling the number of DT they attended and the positiveness of DEI policies. This pattern of results shows that the manipulation could be even effective in reducing unconscious or subconscious biases against racial minorities, but that occurs at the expense of increasing psychological entitlement, which is a conscious psychological state.

In Study 1, although the present manipulations did not trigger moral licensing effects for participants, the results of both direct effects and moderating effects of manipulations strongly demonstrated how effective the manipulations were in lessening DT backlash manifestations. Below, I further discuss how I interpreted the result regarding moral credentials and what were the limitations of Study 1. I also provide how I will address such limitations in Study 2.

As briefly explained in the previous chapters, the moral licensing theory explicates that individuals can acquire moral credentials by conducting good deeds, and doing so often leads to negative outcomes because, with such credentials, how the individuals construe their good deeds and bad deeds are changed (Monin & Miller, 2001). As a result, the theory argues that people with moral credentials engage in immoral deeds without acknowledging the action could be deemed immoral (Effron & Conway, 2015; Merritt et al., 2010). Researchers have empirically investigated the effects of moral credentials and often found that when moral credentials were measured, they were associated with positive behavioral outcomes, such as reducing psychological entitlement and workplace deviance, whereas moral credits were related to increasing psychological entitlement (Loi et al., 2020). Although Loi et al. (2020)'s study used a survey method, such findings show that moral credentials can reduce negative outcomes especially when the moral credentials were measured. However, Lin et al. (2016) also found that obtaining moral credentials increased abusive supervision. These mixed results imply that moral credentials are a theoretically critical construct influencing subsequent behavior, but empirical results may vary. In my study, I initially argued that obtaining moral credentials could increase DT backlash based on the moral licensing theory, but the empirical indications strongly showed that acquiring moral credentials was negatively correlated with a wide range of DT backlash manifestations. Thus, in Study 1, I conclude that increased levels of moral credentials are a state derived from the manipulations that can function as a buffer against DT backlash manifestations.

Besides this issue, there were a few limitations to Study 1. First, the intended effects of manipulations did not elicit moral licensing effects on many DT backlash manifestations. Finding opposite results of the manipulations is reasonable because prior studies confirm training readiness is critical for improving training outcomes (e.g., Hollday et al., 2003; Quinones, 1995). However, if such unexpected results are triggered by how the present manipulations asked participants to recall their previous experiences with DT, it could be problematic. In the experiment, I asked participants multiple questions to recall how their experiences influenced their workplace outcomes and relationships. Researchers have continuously shown that when subjects were given chances to recall their previous moral deeds in an abstract fashion (i.e., considering their good behavior with the superordinate goals of the behavior), they were less likely to engage in moral licensing processes (Conway & Peetz, 2012). On the other hand, when subjects recalled their good deeds with concrete mindsets (i.e., thinking about the good deeds without considering the superordinate goals behind the good deeds), they were more likely to license their previous good deeds (Cornelissen et al., 2013). In Study 2, to effectively trigger the moral licensing processes of participants, I will improve the manipulations by 1) adding a statement for each group fictitiously explaining that DT greatly enhances one's morality, and 2) reducing open-ended questions so that the participants may not fully consider the superordinate goal of their good behavior.

The second limitation is related to the moderators in Study 1. I hypothesized and tested that people's perceptions of unfairness and personality traits would strengthen the relationship between manipulation and DT backlash. However, these moderators, particularly for SDO and BJW had stronger direct effects on DT backlash than manipulations. This result might show that individuals with such personality traits and unfairness perceptions do not necessarily engage in moral licensing processes to express their prejudice or use those credentials in performing behaviors that are against DEI values. Because moral licensing processes involve self-validation of their good deeds (Miller & Effron, 2010), it is plausible that personality traits related to how people view themselves may better predict moral licensing tendencies than personality traits that are related to right-wing authoritarianism, such as SDO and BJW.

Because Study 1 was exploratory in nature, I did not utilize DT for any manipulation groups or control groups. As a result, outcome variables in Study 1 do not fully represent DT backlash manifestations; rather, they are attitudes and behavioral tendencies toward minority group members in general. In Study 2, I assign DT to participants so that the outcome variables represent DT backlash manifestations, and I examine how such outcomes are affected by the manipulations.

STUDY 2

Study 2 also follows a similar procedure to that of Study 1. However, Study 2 employs a short video clip that functions as DT. In addition, I modified manipulations slightly to strengthen the moral licensing processes. Since DT is included in the study, here I also examine how recalling previous DT experiences influences a cognitive manifestation of DT backlash. I also added one more control group in which participants watch DT but do not recall any previous experiences with DT. This was to examine whether participants could be morally licensed by merely participating in DT or watching a DT-related video. Because the results of Study 1 suggest that people high in SDO and BJW are not any more likely to engage in the moral licensing process, I also included other personality traits to further test whether less politically-related personality traits, such as the five-factor personality traits, can predict one's moral licensing processes.

Method

Participants

I also collected samples from Prolific. A total of 1,350 participants were recruited to participate in the screening procedure for Study 2. As in Study 1, the target sample for each group was 100, and Study 2 had three manipulation groups and two control groups. To be eligible for the screening process, the online respondents had to be in North America (US or Cananda) and older than 18 years old. Again, I administered a screening procedure that asked questions about whether they had participated in DT previously. 825 participants have passed the screening procedure. Among 825 participants, 466 participants successfully completed the survey. After removing nine participants who did not pass either manipulation or attention checks, the final number of the sample was 457. Participants received $0.40 for the screening procedure and $8.00 for completing the main experiment.

Procedure

As in Study 1, upon passing the screening procedure, participants received invitations to join the main study. Participants were randomly assigned the subjects to five different groups: three manipulation groups, and one control group. Three manipulations were the same as Study 1. However, to increase the strengths of manipulations and to make participants experience moral licensing processes, I followed the manipulations from Kouchaki (2011) and included statements that fictitiously describe that recent studies have found that people who participated in DT were more likely to make ethical decisions and, as a result, were more moral people. Participants in a group who recalled the positive aspects of their organization's current DEI initiatives read the statement demonstrating that recent studies have found that people in organizations with DEI policies were more moral than companies that do not emphasize such values. After reading those passages, participants were asked to recall previous DT experiences or the positiveness of the DEI policies in the current workplace, depending on their group assignments. Participants in group 1 and 2 were asked to answer multiple open-ended questions, and those questions were related to specific aspects of DT and how DT impacted their behavior in the workplace. Participants in the first manipulation condition were asked to answer how much they liked previous DT and how DT influenced their personal relationships with colleagues from different backgrounds and then watched a DT video. Participants in the second manipulation condition were requested to provide the number of DT they participated in college, at their previous workplace, at their current workplace, and the number of DT they have taken in total. After answering these questions, they watched the DT video. Participants in the third manipulation condition first provided the ongoing DEI policies at their current workplace and then shared how such policies provided a chance to maintain equal opportunities in the workplace and interact better with other colleagues. After answering these questions, they watched the DT video. Participants in control condition 1 were not asked to recall anything but watched the DT video. For control group 2, participants did not recall anything, and they did not watch the DT video. Participants in every condition first performed implicit association tests (IAT) and then went through justice perceptions questionnaires previous to manipulation. After the manipulations, participants watched the DT video and then performed the second round of IATs. After this post-IAT, they answered survey questionnaires related to DT backlash manifestations and demographic questionnaires. Participants were fully debriefed about why they were asked to recall previous DT experiences and were dismissed.

Measures & materials

Participants rated the extent to which they agreed with the statements on a 5-point Likert scale otherwise specified. Measures that were added to Study 2 can also be found in Appendix E. The measures below only include the scales that were newly introduced for Study 2. Internal consistency coefficients can be found in Table in Appendix J.

Diversity Training. I utilized a four-minute video clip to function as a short version of DT. Every treatment group and control group 1 except for control group 2 watched the short clip. I chose such a short video to increase the likelihood of participants concentrating on the video while maintaining some levels of core messages of DEI values that can morally license participants.

DT backlash cognitions – metacognitive activity. To measure the extent to which the participants engage in metacognitive activity during DT, I used an adapted version of the 15-item scale developed by Schmidt and Ford (2003). Because DT applied in Study 2 was very short in duration, participants' ability and effort to engage in metacognitive activities could have been limited. Thus, I only retained five items that were relevant for short training. Those items were "I thought about skills I needed to perform what the video was emphasizing," "I tried to monitor closely the areas where I needed the most improvement," "I thought about what things I needed to do to learn more," "I tried to make sure I understood the things from the video," and "I tried to think through the topic and decide what I was supposed to learn from it."

Control variables. I controlled for social desirability as people with high levels of social desirability may differently answer questionnaires related to DT backlash manifestations.

Analysis

Because the experimental designs I used in Study 1 and 2 were identical except for the additional control group and watching a DT video, the analytical strategy was exactly the same. That is, I conducted ANOVA and ANCOVA to compare the group mean differences of DT backlash between the treatment groups and the control group to examine the main effects of the manipulations. To perform moderator analyses, I dummy-coded each manipulation group, and the referent group for the regression analysis was the control group 2, which did not recall anything and did not watch a DT video. I used multiple regression and hierarchical regression to test the product terms of categorical and continuous variables. As in Study 1, because the results of moderator analyses were based on the dummy variables, unstandardized regression coefficients represent the mean differences from those of the referent group in terms of DVs, and intercepts represent the means of the referent groups' DVs (Cohen et al., 2013). I mean-centered all the moderators, and I did not center the independent variables because they were dummy coded variables.

RESULTS

Table 13 in Appendix J shows descriptive statistics, Cronbach's alpha for each variable, and correlation coefficients for Study 2, respectively. Manipulation group 1 had 89 participants, manipulation group 2 had 93 participants, manipulation group 3 had 86 participants, control group 1 had 90 participants, and control group 2 had 99 participants.

Main Effects of Manipulations on DT Backlash

The results of one-way ANOVA showed that the effects of manipulations were statistically significant on some DT backlash variables, including moral credentials ($F(1, 457) = 3.20$, $p < .05$, $\eta^2 = .03$; (manipulation group 1 *M* = 4.03 *SD* = 0.83, manipulation group 2 *M* = 3.75 *SD* = 1.05, manipulation group 3 *M* = 4.17 *SD* = 0.87, control group 1 *M* = 4.04, *SD* = 0.83; control group 2 *M* = 3.82 *SD* = 1.03), and metacognitive activity ($F(1, 457) = 4.37$, $p < .01$, $\eta^2 = .04$; manipulation group 1 *M* = 4.13 *SD* = 0.65, manipulation group 2 *M* = 3.77 *SD* = 0.82, manipulation group 3 *M* = 4.04 *SD* = 0.70, control group 1 *M* = 3.98, *SD* = 0.81; control group 2 *M* = 3.75 *SD* = 0.84), but the direction of the effects was opposite from Hypothesis 1a, 1b, and 1c, failing to support these hypotheses. Figures 23 to 26 in Appendix K summarize the direct effects of each group on DT backlash manifestations. For moral credentials, Tukey's HSD post hoc analysis showed that the mean difference between manipulation group 2 (*M* = 3.75, *SD* = 1.05), and manipulation group 3 (*M* = 4.17, *SD* = 0.87) was statistically significant (*M* difference = -.42, SE = .14, $p < .05$). Also, the mean difference between manipulation group 3 (*M* = 4.17, *SD* = 0.87) and control group 2 (*M* = 3.82, *SD* = 0.94) was also significant at $p < .10$ level (*M* difference = .35, SE = .14, $p < .10$). For metacognitive activity, the post hoc analysis showed that the mean difference between manipulation group 1 (*M* = 4.13, *SD* = 1.05) who recalled the previous DT experiences and manipulation group 2 (*M* = 3.77, *SD* = 0.87) who recalled the number of previous DT was statistically significant (*M* difference = .36, SE = .11, $p < .05$). In addition, the mean of manipulation group 1 (*M* = 4.13, *SD* = 1.05) was statistically different from that of the control group 2 who did not recall anything nor watched a DT video (*M* = 3.83, *SD* = .84; *M* difference = .38, SE = .11, $p < .01$). Lastly, the mean of manipulation group 3 (*M* = 4.04, *SD* = .70) and that of control group 2 (*M* = 3.83, *SD* = .84) were significantly different at $p < .10$ level (*M* difference = .28, SE = .11) in terms of metacognitive activities.

The effects of manipulations on hiring decisions were significant at $p < .10$ level ($F(1, 457) = 1.91$, $\eta^2 = .02$; manipulation group 1 *M* = 4.10 *SD* = 0.99, manipulation group 2 *M* = 3.83 *SD* = 1.23, manipulation group 3 *M* = 4.02 *SD* = 1.10, control group 1 *M* = 4.29, *SD* = 1.32; control group 2 *M* = 4.15 *SD* = 1.26). Participants in manipulation group 2 (*M* = 3.83, *SD* = 1.23) who recalled the number of previous DT showed lower levels of hiring majority group members compared to participants in control group 1 *M* = 4.29, *SD* = 1.32; *M difference* = -.46, SE = .18, $p < .10$). As in Study 1, higher scores in hiring decisions indicate that respondents tend to prefer majority group members to minority group members for the job.

Among discrete emotions, ANOVA results showed that group mean differences were significant among all six emotions. Group means were statistically different for anger [$F(1, 457) = 2.74$, $p < .05$, $\eta^2 = .02$], fear [$F(1, 457) = 2.03$, $p < .10$, $\eta^2 = .02$], joy [$F(1, 457) = 10.40$, $p < .001$, $\eta^2 = .08$], love [$F(1, 457) = 11.61$, $p < .001$, $\eta^2 = .09$], sadness [$F(1, 457) = 2.30$, $p < .10$, $\eta^2 = .02$], and surprise [$F(1, 457) = 2.37$, $p < .10$, $\eta^2 = .02$]. Because anger is the primary emotion felt when one is experiencing DT backlash, I only report the results of moderation analyses for anger (manipulation group 1 *M* = 1.24 *SD* = 0.72, manipulation group 2 *M* = 1.41 *SD* = 0.93, manipulation group 3 *M* = 1.17 *SD* = 0.60, control group 1 *M* = 1.24, *SD* = 0.65; control group 2 *M* = 1.48 *SD* = 0.84).

I also ran regression analyses to test the significance of mean differences between each group using the *t* test. Table 14 in Appendix K summarizes the result. The referent group was control group 2. Results of the regression analyses on moral credentials demonstrated that the mean differences between manipulation group 3 (*b* = .35, *SE* = .14, *t* = 2.50, *p* < .05) and the control group 2 (*intercept* = 3.82, *SE* = .09, *t* = 42.44, *p* < .001). Control group 1 also demonstrated the mean differences with control group 2 at *p* < .10 level (*b* = .22, *SE* = .14, *t* = 1.57, *p* < .05). For metacognitive activity, manipulation group 1 (*b* = .38, *SE* = .11, *t* = 3.45, *p* < .01), manipulation group 3 (*b* = .29, *SE* = .11, *t* = 2.64, *p* < .05), and control group 1 (*b* = .23, *SE* = .11, *t* = 2.09, *p* < .05) showed mean differences with control group 2 (*intercept* = 3.76, *SE* = .09, *t* = 41.78, *p* < .001). For explicit prejudice, manipulation group 1 (*b* = -.19, *SE* = .10, *t* = 1.90, *p* < .10), manipulation group 3 (*b* = -.21, *SE* = .10, *t* = 2.10, *p* < .05) showed mean differences with control group 2 (*intercept* = 2.13, *SE* = .07, *t* = 30.43, *p* < .001), but the mean difference between manipulation group 1 and control group was significant at *p* < .10 level.

For OCB, manipulation group 3 (*b* = .18, *SE* = .11, *t* = 1.64) showed mean differences with control group 2 (*intercept* = 4.07, *SE* = .07, *t* = 58.14, *p* < .001) at *p* < .10 level. For hiring decision, manipulation group 2 (*b* = -.32, *SE* = .17, *t* = 1.88) showed mean differences with control group 2 (*intercept* = 4.15, *SE* = .12, *t* = 34.58, *p* < .001) at *p* < .10 level.

Moderating Effects of Distributive Justice Perceptions

Table 15 in Appendix L and Figures 27 to 33 summarize the moderation analysis results for distributive justice. According to the analyses, distributive justice perceptions had direct effects on anger (*b* = -.27, *SE* = .03, *t* = -9.11, *p* < .001), moral credentials (*b* = .56, *SE* = .03, *t* = 19.88, *p* < .001), moral credits (*b* = .48, *SE* = .09, *t* = 5.70, *p* < .001), metacognitive activity, explicit prejudice (*b* = -.32, *SE* = .06, *t* = -5.69, *p* < .001), and OCB (*b* = .22, *SE* = .06, *t* = 3.87, *p*

$< .001$). The moderating effects of distributive justice were significant for anger, moral credentials, and metacognitive activity. Hypothesis 2 was concerned with the moderating effect of distributive justice on the relation between manipulations and DT backlash. The product terms of manipulation groups and distributive justice were statistically significant for manipulation group 1 ($b = .34$, $SE = .09$, $t = 3.74$, $p < .01$), manipulation group 3 ($b = .20$, $SE = .09$, $t = 2.07$, $p < .05$), and control group 1 ($b = .33$, $SE = .09$, $t = 3.59$, $p < .01$) on anger. A simple slope analysis showed that manipulation 1 had a negative relation with anger when participant's distributive justice was low ($b = -.67$, $SE = .14$, $t = -4.63$, $p < .001$), while not significant for participants with high distributive justice ($b = .11$, $t = 0.73$, *ns*); manipulation 3 had a negative relation with anger when participant's distributive justice was low ($b = -.52$, $SE = .16$, $t = -3.31$, $p < .01$), while not significant for participants with high distributive justice ($b = -.07$, $t = -0.51$, *ns*); control group 1 also had a negative relation with moral credentials when participant's distributive justice was low ($b = -.65$, $SE = .15$, $t = -4.29$, $p < .001$), while not significant for participants with high distributive justice ($b = .09$, $t = 0.67$, *ns*).

The product terms of manipulation groups and distributive justice were statistically significant for manipulation group 1 ($b = -.29$, $SE = .09$, $t = -3.29$, $p < .01$), manipulation group 3 ($b = -.20$, $SE = .09$, $t = -2.20$, $p < .05$), and control group 1 ($b = -.18$, $SE = .09$, $t = -2.01$, $p < .05$) on moral credentials. A simple slope analysis showed that manipulation 1 had a positive relation with moral credentials when participant's distributive justice was low ($b = .61$, $SE = .14$, $t = 4.32$, $p < .001$), while not significant for participants with high distributive justice ($b = -.06$, $t = -0.41$, *ns*); manipulation 3 had a positive relation with moral credentials when participant's distributive justice was low ($b = .55$, $SE = .16$, $t = 3.54$, $p < .001$), while not significant for participants with high distributive justice ($b = .08$, $t = 0.58$, *ns*); control group 1 also had a positive relation with moral credentials when participant's distributive justice was low (*b* = .49, *SE* = .15, *t* = 3.36, *p* < .01), while not significant for participants with high distributive justice (*b* = .08, *t* = 0.58, *ns*).

For metacognitive activity, the product term of the manipulation group 1 and distributive justice was significant (*b* = -.15, *SE* = .09, *t* = -1.68) at *p* < .10 level. A simple slope analysis showed that manipulation 1 had a positive relation with moral credentials when participant's distributive justice was low (*b* = .59, *SE* = .04, *t* =13.59, *p* < .001), while not significant for participants with high distributive justice (*b* = .24, *t* = 1.23, *ns*).

Because the interaction between manipulations and distributive justice showed that manipulations lessened DT backlash for participants with low distributive justice perceptions on anger, moral credentials, and metacognitive activity, Hypothesis 2 was not supported.

Moderating Effects of Procedural Justice Perceptions

Table 16 in Appendix L and Figures 34 to 37 summarize the moderation analysis results for procedural justice. In Study 2, I measured only one facet of procedural justice as the two facets measured in Study 1 were highly correlated (*r* = .89). Thus, I only measured the extent to which trainees believe that their assignment to DT was fair in terms of procedural justice in Study 2, and it had direct effects on anger (*b* = -.30, *SE* = .04, *t* = -8.12, *p* < .001), moral credentials (*b* = .49, *SE* = .04, *t* =11.82, *p* < .001), moral credits (*b* = .53, *SE* = .05, *t* = 10.72, *p* < .001), metacognitive activity (*b* = .36, *SE* = .04, *t* =9.94, *p* < .001), explicit prejudice (*b* = -.16, *SE* = .03, *t* = -4.84, *p* < .001), and OCB (*b* = .25, *SE* = .04, *t* = 7.30, *p* < .001). Hypotheses 3 was concerned with the moderating effect of procedural justice perceptions on the relationships between manipulations and DT backlash. The moderating effects of procedural justice on the relationships between manipulations and DT backlash were not statistically significant except for anger. For anger, the product terms of manipulation groups and procedural justice were statistically significant for manipulation group 1 (*b* = .28, *SE* = .11, *t* = 2.44, *p* < .05), manipulation group 3 (*b* = .23, *SE* = .11, *t* = 2.03, *p* < .05), and control group 1 (*b* = .36, *SE* = .11, *t* = 3.27, *p* < .01) on anger. A simple slope analysis showed that manipulation 1 had a negative relation with anger when participant's procedural justice was low (*b* = -.47, *SE* = .15, *t* = -3.15, *p* < .01), while not significant for participants with high procedural justice (*b* = .05, *t* = 0.33, *ns*); manipulation 3 had a negative relation with anger when participant's procedural justice was low (*b* = -.45, *SE* = .16, *t* = -2.88, *p* < .01), while not significant for participants with high procedural justice (*b* = -.02, *t* = -0.12, *ns*); control group 1 also had a negative relation with anger when participant's procedural justice was low (*b* = -.56, *SE* = .15, *t* = -3.83, *p* < .001), while not significant for participants with high procedural justice (*b* = .12, *t* = 0.82, *ns*).

Because the interaction between manipulations and procedural justice showed that manipulations lessened DT backlash for participants with low procedural justice perceptions on anger, Hypothesis 3 was not supported.

Moderating Effects of Personality Traits

I also conducted regression analyses to test moderating effects of SDO and BJW between the manipulations and DT backlash. Table 17 in Appendix L and Figures 38 to 43 summarize the moderation analysis results for SDO. SDO had direct effects on implicit prejudice (*b* = -2.25, *SE* = .08, *t* = -2.92, *p* < .01), anger (*b* = .44, *SE* = .05, *t* = 8.77, *p* < .001), moral credentials (*b* = -.81, *SE* = .05, *t* = -14.97, *p* < .001), moral credits (*b* = -.30, *SE* = .08, *t* = -3.96, *p* < .001), psychological entitlement (*b* = .23, *SE* = .05, *t* = 4.78, *p* < .001), metacognitive activity (*b* = -.35, *SE* = .05, *t* = -6.66, *p* < .001), explicit prejudice (*b* = .68, *SE* = .04, *t* = 18.83, *p* < .001), OCB (*b* = -.55, *SE* = .04, *t* = -12.54, *p* < .001), and willingness to hire minority group members (*b* = .17, *SE* = .08, *t* = 1.99, *p* < .05). Hypothesis 4 proposed that the interaction between manipulations and

SDO would strengthen DT backlash as SDO increases. The moderating effects of SDO were statistically significant on DT backlash for anger, metacognitive activity, explicit prejudice, and OCB. For anger, the product term of manipulation group 1 and SDO on anger (*b* = -.29, *SE* = .16, *t* = -1.83) was statistically significant at *p* < .10 level. A simple slope analysis showed that manipulation 1 had a negative relation with anger when participant's SDO was high (*b* = -.38, *SE* = .15, *t* = -2.50, *p* < .01), while not significant for participants with low SDO (*b* = .01, *t* = 0.04, *ns*).

For metacognitive activity, the product term of manipulation group 1 and SDO on metacognitive activity (*b* = .30, *SE* = .17, *t* = 1.76) was statistically significant at *p* < .10 level. A simple slope analysis showed that manipulation 1 had a positive relation with metacognitive activity when the participant's SDO was high (*b* = .55, *SE* = .16, *t* = 3.46, *p* < .01), while not significant for participants with low SDO (*b* = .16, *t* = 1.07, *ns*).

For explicit prejudice, the product terms of manipulation group 2 and SDO (*b* = -.25, *SE* = .10, *t* = -2.39, *p* < .05), and manipulation group 3 and SDO (*b* = -.25, *SE* = .11, *t* = -2.34, *p* < .05) on explicit prejudice were statistically significant. A simple slope analysis showed that manipulation 2 had a negative relation with explicit prejudice when participant's SDO was high (*b* = -.24, *SE* = .11, *t* = -2.28, *p* < .05), while not significant for participants with low SDO (*b* = .09, *t* = 0.92, *ns*); manipulation 3 had a negative relation with explicit prejudice when participant's SDO was high (*b* = -.32, *SE* = .10, *t* = -3.07, *p* < .01), while not significant for participants with low SDO (*b* = .02, *t* = 0.19, *ns*).

For the willingness to help minority group members (OCB toward minorities), only the product term between manipulation group 3 and SDO (*b* = .27, *SE* = .13, *t* = 2.03, *p* < .05) was significant. A simple slope analysis demonstrated that manipulation 3 had a positive relation with

OCB when participant's SDO was high (*b* = .31, *SE* = .13, *t* = 2.48, *p* < .05), while not significant for participants with low SDO (*b* = -.04, *t* = -0.32, *ns*). As the results showed that interactions between the manipulations and SDO reduced DT backlash for individuals with high SDO, Hypothesis 4 was not supported.

Additionally, the interaction between control group 1 and SDO showed that control group 1 had a negative relation with implicit prejudice when participant's SDO was high (*b* = -6.53, *SE* = 3.46, *t* = -1.89) at *p* < .10 level, while not significant for participants with low SDO (*b* = 3.37, *t* = 0.92, *ns*). Also, the interaction between control group 1 and SDO on moral credits was significant at *p* < .10 level (*b* = -.38, *SE* = .23, *t* = -1.63), while both simple slopes for SDO were not significant.

Table 18 in Appendix L and Figures 44 to 47 summarize the moderation analysis results for BJW. BJW had direct effects on moral credentials (*b* = -.25, *SE* = .07, *t* = -3.68, *p* < .001), moral credits (*b* = .35, *SE* = .08, *t* = -4.50, *p* < .001), psychological entitlement (*b* = .19, *SE* = .05, *t* = 3.94, *p* < .001), explicit prejudice (*b* = .37, *SE* = .05, *t* = 8.06, *p* < .001), OCB (*b* = -.20, *SE* = .05, *t* = -3.82, *p* < .001), and willingness to hire minority group members (*b* = .20, *SE* = .09, *t* = 2.37, *p* < .05). The moderating effects of BJW were statistically significant for moral credentials, psychological entitlement, and hiring decisions. The product term of manipulation group 1 and BJW (*b* = .47, *SE* = .21, *t* = 2.28, *p* < .05) was significant on moral credentials. The interaction between control group 1 and BJW was significant at *p* < .10 level (*b* = .35, *SE* = .20, *t* = 1.73). A simple slope analysis showed that manipulation 1 had a positive relation with moral credentials when participant's BJW was high (*b* = .55, *SE* = .20, *t* = 2.76, *p* < .01), while not significant for participants with low BJW (*b* = 0.06, *t* = -0.34, *ns*); control group 1 had a positive relation with moral credentials when participant's BJW was high (*b* = .49, *SE*

= .18, *t* = 2.65, *p* < .01), while not significant for participants with low BJW (*b* = 0.03, *t* = 0.17, *ns*).

For psychological entitlement, the product term of manipulation group 2 and BJW (*b* = .28, *SE* = .15, *t* = 1.85) was significant at *p* < .10 level. A simple slope analysis revealed that manipulation 2 had a negative relation with psychological entitlement when participant's BJW was low (*b* = -.23, *SE* = .13, *t* = -1.73) at *p* < .10 level, while not significant for participants with high BJW (*b* = 0.13, *t* = 0.93, *ns*).

For hiring decisions, the product term of manipulation group 2 and BJW (*b* = .47, *SE* = .27, *t* = 1.75) was statistically significant on hiring decisions. A simple slope analysis showed that manipulation group 2 had a negative relation with hiring decisions when participant's BJW was low (*b* = -.63, *SE* = .24, *t* = -2.65, *p* < .01), while not significant for participants with high BJW (*b* = -0.02, *t* = -0.08, *ns*).

For metacognitive activity, the interaction between control group 1 and BJW was significant at p < .10 level (*b* = .33, *SE* = .17, *t* = 1.90). A simple slope analysis revealed that control group 1 had a positive relation with metacognitive activity when participant's BJW was high (*b* = .41, *SE* = .15, *t* = 2.68, *p* < .01), while not significant for participants with low BJW (*b* = -0.01, *t* = -0.06, *ns*).

The results showed that interactions between manipulations and BJW reduced DT backlash, particularly for participants with high BJW. Thus, Hypothesis 5 was not supported.

Supplemental Analyses

As in Study 1, I conducted additional analyses that controlled for psychological reactance to test whether the relations among hypothesized variables hold after adding such personality traits. The results show that all of the results for the moderation analyses remained unchanged, while there were some of changes to the direct effects of manipulations. Before controlling psychological reactance, there were self-consistency effects of manipulations on explicit prejudice for manipulation groups 2 and 3, on OCB for manipulation group 3, and on moral credentials for control group 1. However, these self-consistency effects were nullified after adding psychological reactance.

Similar results were obtained when I controlled for social desirability. All of the results for moderation analyses remained unchanged, but two of the direct effects of manipulations were gone after the introduction of the control. Those were the effect of the manipulation group 3 on OCB and the effect of control group 1 on moral credentials.

Discussion

Table 20 in Appendix N summarizes all the significant results of Study 2. Similar to the findings from Study 1, the results from the statistical analyses demonstrate that the manipulations generated self-consistency effects rather than moral licensing effects. In spite of my effort to strengthen moral licensing processes by manipulations in Study 2, the statistical results showed that the main effects of recalling previous experiences of DT or the positiveness of the current DEI policies lessened DT backlash. As in Study 1, the manipulations were very effective at reducing DT backlash manifestations, such as metacognitive activity and explicit prejudice. Recalling previous experiences with DT not only increased one’s level of metacognitive activity when watching a DT video but also significantly reduced explicit prejudice compared to the control group. Recalling the number of DTs that participants attended led to a higher likelihood of hiring minority group members. Recalling the positiveness of DEI policies at their current workplace also increased moral credentials, metacognitive activity during DT, and willingness to help minority group members while reducing explicit prejudice. Although some of the results were significant at $p < .10$ level, the manipulations, in general, decreased the extent to which the participants expressed DT backlash. Interestingly, one of the control groups in which participants did not recall anything but watched the DT video reported an increase in moral credentials and metacognitive activity compared to the other control group, suggesting that engaging in DT without thinking about previous DT can still generate positive outcomes. Taken together, the results from Study 2 replicate Study 1 and more clearly demonstrate that recalling previous DT is a powerful psychological mechanism that can diminish DT backlash.

As in Study 1, a moderation analysis of justice perceptions demonstrates that manipulations were highly effective for those who had lower distributive justice perceptions and procedural perceptions. When participants with low distributive justice recalled their past experience of DT, doing so increased moral credentials and metacognitive activity while decreasing anger. Also, when participants with low distributive justice recalled the positiveness of the current DEI policies, they were more likely to acquire moral credentials while decreasing anger. For participants with low distributive justice, watching the DT video without recalling anything helped the participants obtain moral credentials and reduce anger.

In a similar manner, interactions between manipulations and procedural justice perceptions generated consistency effects. When participants with low procedural justice recalled either their past experience of DT or the positiveness of the current DEI policies, it reduced participants' anger. Again, these results suggest that the manipulations were more effective for participants with unfairness perceptions.

The moderation analyses of personality traits also demonstrated that manipulations used in Study 2 tend to lessen the effects of DT backlash variables. In Study 2, the effects of interactions between SDO and manipulations become more pronounced. When participants with high SDO recalled past experience of DT, they were more likely to reduce anger, while increasing metacognitive activity. For those participants, recalling the number of DT that they attended also led to a reduction of explicit prejudice. Additionally, recalling the positiveness of the current DEI policies for participants with high SDO led to a reduction of explicit prejudice and an increase in willingness to help racial minorities. For participants with high SDO, watching the DT video without recalling anything helped the participants decrease their implicit prejudice.

With regard to the interaction between manipulations and BJW, BJW's moderation effects were less pronounced than in Study 1. When participants with high BJW recalled previous experiences of DT, it led to higher levels of moral credentials. BWJ did not interact with other manipulation groups. However, it interacted with a control group whose participants watched the DT video but did not recall anything to produce higher levels of moral credentials and metacognitive activity for high BJW participants. This result suggests that the manipulations were, again, effective at lessening DT backlash for participants who have stable personality traits that are related to increased prejudice against social minorities.

The findings from Study 2 generally replicated the findings from Study 1 that the manipulations used in both studies were powerful drivers to reduce DT backlash manifestations, and the moderation analyses further confirmed that recalling various aspects of DT can be critical for improving trainees who have lower levels of fairness perceptions or have higher levels of personality traits that are prone to prejudice. Study 2 replicated such results after strengthening the manipulations to arouse participants' moral licensing processes. Thus, I argue that recalling various aspects of DT does not provide opportunities for participants to morally license themselves, but it can motivate trainees to engage better in subsequent DT. As such results highlight there are crucial theoretical and practical implications, I will discuss such implications in the next section. I also address the limitations of Study 2 and future research avenues that can shed light on DT backlash.

CHAPTER 4: GENERAL DISCUSSION

This dissertation aimed to advance the understanding of the DT backlash phenomenon by developing its conceptualization as a scientific construct and empirically testing specific theory-based predictions regarding the role of moral credentialing in DT backlash. As a first step to conceptualizing DT backlash, I systematically reviewed the DT backlash literature, identifying the different ways in which DT backlash has been defined and examining the various theoretical lenses researchers have utilized to attempt to explain DT backlash. I provided a systematic and critical review of the literature that highlighted that 1) there is no consensus in DT backlash definitions, and 2) the literature suffers from the accumulation of studies that are based on the inconsistent and often deficient definitions of DT backlash. Conducting such a review contributes to the DT backlash literature because identifying the inconsistencies and limitations of current definitions and developing a conceptually sound and widely agreed-upon definition of DT backlash is an important step in advancing the understanding of the phenomenon.

To address the current limitations in the literature, I proposed a comprehensive definition of DT backlash that captures key elements of the phenomenon, such as how DT backlash manifests itself, the dimensions of such manifestations, and who is more likely to experience it. My proposed definition also incorporates fundamental and unique characteristics of DT backlash. For example, I described the differences between DT backlash and DT effectiveness and the potential targets of DT backlash. By doing so, this dissertation paves the way for future researchers to utilize the definition I proposed and build literature based on a shared definition that describes the core elements of DT backlash. This is important for the relatively scant literature on DT backlash because accumulating empirical studies with a consensual definition would help researchers understand the findings and interpret effect sizes.

In addition to revealing the lack of a consensus definition of the DT backlash construct, my critical review also revealed that the current literature's understanding of the potentially significant impact the moral licensing process may have on DT backlash is greatly underdeveloped. Researchers have identified that moral licensing processes could be one of the psychological mechanisms that drive DT backlash (e.g., Leslie, 2019), but the moral licensing effects on DT backlash have not been empirically tested in the literature. To theoretically and empirically address this issue, I proposed hypotheses and tested whether one's moral licensing process can elicit DT backlash. Specifically, I tested whether recalling past experiences of DT, the number of DT that one participated in, or the positiveness of DEI policies in the current workplace can become moral licenses or credentials for research participants, which, in turn, triggers DT backlash. To further examine the boundary conditions of such processes on DT backlash, the empirical tests investigated the role of trainee characteristics: justice perceptions and personality traits such as SDO and BJW. I designed and conducted two online experiments on Prolific and collected over 800 participants to test my hypotheses. Developing hypotheses and testing such ideas contribute to the DT backlash literature by 1) linking a unique perspective from the moral licensing theory to DT backlash and 2) empirically testing the effects of moral licensing influences DT backlash using rigorous experimental designs.

Overall, the results of both Study 1 and Study 2 unexpectedly showed that the manipulations created the opposite effects from the moral licensing effects. Although I found the effects of recalling the DT-related experiences on DT backlash that were opposite from what I hypothesized, my findings convey important information about how recalling DT-related experiences can reduce DT backlash. That is, participants who experienced the manipulations were more likely to lessen DT backlash compared to the control group(s) in both studies. The results further show that recalling prior experiences of DT or the positiveness of DEI policies was particularly effective and beneficial for reducing cognitive, emotional, and affective manifestations of DT backlash, while such effects were weaker for behavioral manifestations. These results highlight that recalling DT-related experiences generated self-consistency effects rather than moral licensing effects, and they were highly effective in reducing DT backlash.

Another unique aspect of these results regarding the direct effects of the manipulations is that recalling DT experiences without DT reduced one's explicit prejudice and anger (Study 1), and such effects were also found after they watched a short version of DT (Study 2). This further shows that recalling DT-related experiences can enhance one's pretraining attitudes, reducing DT backlash. Also, the results from Study 2 revealed that the responses of participants who recalled such experiences demonstrated less evidence of DT backlash (e.g., moral credentials, metacognitive activity) in a greater magnitude than participants who watched the DT video without recalling anything (control group 1). Also, recalling DT-related experiences lessened DT backlash from cognitive to behavioral dimensions, while the group that experienced DT without recalling anything showed changes in cognitive manifestations. Taken together, recalling past experiences of DT is an influential and beneficial psychological process that leads to a reduction in DT backlash and, perhaps, an increase in DT effectiveness.

The results of moderating effects of justice perceptions and personality traits provide an even more compelling picture of the effects of manipulations on DT backlash. I predicted that participants with low justice perceptions and high SDO or BJW would increase DT backlash when they recall their previous experience with DT because they would be more likely to use their experience as a moral license to exert their feelings of unfairness or deep-seated prejudice. However, findings again revealed that the manipulations reduced DT backlash, and such effects were stronger for participants with low justice perceptions, high SDO, or high BJW. This tendency occurred while unfairness perceptions, SDO, and BJW were directly and positively associated with DT backlash, highlighting that recalling DT-related experiences countered those variables' direct positive effects on DT backlash. This further implies that simply reliving the previous DT can have meaningful effects on particular trainees and also shows that recalling prior DT experiences could be used as a motivation strategy for practitioners. This will be further discussed in the practical implication section.

Results of the simple slope analyses further indicate that participants who perceive DEI initiatives and the assignment of DT as fair, who have low levels of SDO, or who has low levels of BJW tend to show low levels of DT backlash manifestations. And, more importantly, the manipulation conditions did not change such low levels of DT backlash for those participants. That is, such participants in both manipulation groups and a control group showed significantly lower levels of DT backlash manifestations than those who had lower levels of justice perceptions or higher levels of SDO and BJW. Theoretically, it is also plausible that people with high levels of justice perceptions and low levels of SDO or BJW might be more likely to use their positive traits as moral licenses because recalling positive traits of themselves often leads to moral licensing processes (Blanken et al., 2014; Conway & Peetz, 2012; Sachdeva et al., 2009). However, the results of my dissertation demonstrate that people with such positive characteristics do not tend to engage in moral licensing processes. Still, they tend to maintain low levels of DT backlash regardless of recalling DT-related experiences.

Given that my manipulations followed the prior studies that experimentally examined moral licensing (Blanken et al., 2012, Bradley-Geist et al., 2010; Kouchaki et al., 2018) and are deliberately designed to elicit moral licensing effects, finding opposite effects of the manipulations on DT backlash strongly suggests that, in the DT context, recalling past experiences of DT motivate trainees rather than making them licensed about their previous experiences. Also, such tendencies were replicated in Study 2 where I attempted to strengthen the manipulations to trigger participant's moral licensing processes. Taken together, findings from both Study 1 and Study 2 strongly suggest that participants' ability to link their experience with DT to the current DT can increase their pretraining motivation and training readiness, which will further reduce DT backlash manifestations and increase DT effectiveness. Also, the results particularly emphasize that the beneficial effects of recalling DT-related experiences were more pronounced for participants who have high levels of SDO, BJW, and low levels of justice perceptions.

Between Studies 1 and 2, the effects of manipulations were stronger in eliciting self-consistency effects on moral credentials and explicit prejudice in Study 1, but the effects of manipulations were stronger for metacognitive activity and anger in Study 2. These different results might be attributed to 1) the differences in the method used to manipulate participants or 2) whether the study involved DT or not. Participants in Study 1 were asked to recall and write down their experiences and then answered multiple questionnaires. On the other hand, participants in Study 2 read statements that described those who participated in DT were more moral and were better able to make moral decisions than those who did not take DT. After reading the statement, they were asked to recall and write down DT-related experiences and, then watched a short video. Participants in Study 2 were asked to recall fewer DT experiences than participants in Study 1 in order to strengthen the moral licensing processes. Hence, the manipulations and experimental design in Study 1 test how recalling DT-related experiences influences participants' changes in their general attitudes and behavior toward minority group members, while the experimental design in Study 2 examines training outcomes, which include attitudes and behaviors. Perhaps the participants in Study 2 experienced changes in moral credentials or explicit prejudice like participants in Study 1, but such changes may not have lasted after they participated in a short version of DT.

Also, Study 1 highlighted the beneficial effects of the manipulations on trainees with high BJW and less emphasized effects for trainees with high SDO. However, in Study 2, trainees with high SDO were the ones who experienced higher self-consistency effects across various DT backlash manifestations, whereas such effects were less observed for trainees with BJW. The differences in the results further imply that manipulations and DT itself may have interacted to influence participants in both studies differently. The current results describe that the combination of recalling DT-related experiences and DT participation can reduce DT backlash manifestations in various dimensions, but not for trainees high in BJW.

Although the results did not support my hypotheses based on moral licensing theory, the findings of my dissertation are consistent with a stream of research in the literature from both traditional training and DT. Researchers have found that trainees' perceptions of organizational characteristics or DT characteristics influence pretraining motivation, which, in turn, affects training outcomes (e.g., Facteau et al., 1995). Also, Quinones (1995) has found that training assignments that can offer feedback about trainees' past training performance influence training motivation and training performance. Moreover, the training transfer literature explicates that maintenance of training transfer that resulted from the previous training can be critical in training transfer for subsequent training (Ford & Kraiger, 1995), implying that one's previous training experiences can affect training outcomes. In the DT context, Holladay and colleagues (2003) found that trainees' pretraining perceptions of training characteristics can influence trainees' performance in DT. In a similar vein, the results of my dissertation also demonstrate that thinking about previous DT experiences can change one's attitudes and motivation, which was also effective at reducing DT backlash.

However, there were also minor empirical indications that moral licensing processes have occurred in my studies. In Study 1, participants with high SDO reduced their implicit biases after recalling the number of DT that they attended at the expense of heightened psychological entitlement, which is a state that is often increased by peoples' previous good deeds (e.g., Loi et al., 2020; Yam et al., 2017). However, the manipulations did not affect any DT backlash manifestations other than implicit prejudice. This effect was absent in Study 2 where I strengthened the fidelity of the manipulations. Additionally, when participants with high procedural justice recalled the positiveness of DEI policies at their current workplace, they reported that they were less willing to help racial minorities. This result demonstrates that the manipulations could influence behavioral manifestations of DT backlash while not inducing other manifestations. This effect was also not present in Study 2. However, it still suggests that recalling DT-related experiences may elicit moral licensing effects on specific manifestations of DT backlash, whereas the overall results describe the manipulations' strengths to elicit self-consistency effects that were more prevalent and occurred on various manifestations. Lastly, the effects of manipulations on the dependent variables are rather transient in nature, and as a result, their effects are not durable to be observed after the short DT has taken place.

Theoretical Implications

This dissertation offers several theoretical insights into DT backlash and moral licensing theory. I provided a systematic review of DT backlash literature and identified the commonly studied psychological mechanisms of DT backlash and the underdeveloped area. To address the underdevelopment, this study focused on why and how DT backlash occurs. I addressed the limitation regarding the proliferated definitions in the current literature and provided a definition that clarifies and specifies what the backlash is, how it can manifest itself in different reactions, and what would be the unit of analysis. To generate new insight about what may trigger DT backlash, I draw on moral licensing theory to predict how recalling one's past DT experiences may induce DT backlash. By doing so, this dissertation not only utilizes the proposed definition of DT backlash and shows how it can be empirically measured and tested but also expands the current knowledge by inviting and testing the psychological mechanism that is important in understanding DT backlash yet has received little attention.

One significant theoretical implication of this part of my dissertation is that it opens up new avenues for applying the DT backlash construct in the DT literature. By theorizing that DT can produce negative effects on several dimensions, I offer a unique perspective that expands the current conceptualization of DT effectiveness. This dissertation provides more depth to the concept of DT effectiveness by incorporating and specifying the group of outcomes that could be negatively affected by taking DT. I argue that such outcomes are critical in further understanding the success and failure of DT. Thus, future research could focus on DT backlash as a core set of DT outcomes. Since most of the studies in the DT literature tend to focus on increasing DT effectiveness, future research may focus on the factors that increase or decrease DT backlash, which will generate important knowledge about managing diversity issues in organizations. Initial speculation would be that an organization that decouples from the DEI initiatives and provides DT without committing to DEI initiatives could increase DT backlash because such decoupling will generate mixed signals for trainees (e.g., Leslie, 2019).

This dissertation conceptualizes DT backlash and reviews the extant literature on DT backlash, but it does not focus on developing a comprehensive framework for DT backlash. However, because the DT backlash research is in its nascent stage, building a comprehensive nomological network of the construct can generate a greater understanding of the phenomenon. The literature identifies multiple psychological mechanisms, such as psychological reactance and justice perceptions, and some training characteristics, such as training assignments and training titles (Holladay et al., 2003), but it still significantly lacks a systematic understanding of which predictors, contextual factors, and trainee characteristics are involved to influence DT backlash. In addition, there should be differences in levels among factors, which are yet to be specified in the literature. It is highly imperative to acknowledge factors at each level to systematically understand DT backlash. For example, future research may explicate how organizational-level factors (e.g., organization's expenditure on DEI values, organizational culture), team-level factors (e.g., team climate toward DEI values, justice climate), and individual-level factors (e.g., personality traits, commitment to DEI values) are related each other in generating DT backlash.

Furthermore, I measured DT backlash using relevant scales from the DT literature and found statistically significant relations between the manipulations and DT backlash after participating in short DT (Study 2). This could shed light on interpreting and understanding of DT effectiveness. DT backlash was conceptualized as it may occur in one of the manifestations when learning in other dimensions was successfully achieved. This means that effective DT may also suffer from some levels of DT backlash, and if one does not measure DT backlash, then researchers may not acknowledge such negative effects have occurred in DT. If future researchers are interested in evaluating the overall effectiveness of DT, they may incorporate the measurements or questionnaires that I used, such as explicit prejudice and IATs, to identify whether DT backlash occurred along with evidence of DT effectiveness. Thus, by using measures, future studies may focus on revealing why sometimes there are mixed.

My dissertation also sheds light on moral licensing theory in the DT setting. The results of Study 1 and Study 2 demonstrate that recalling previous moral experiences in the DT context generated consistency effects instead of moral licensing processes. Additionally, watching short DT videos without recalling anything did not morally license participants as well. This indicates that in the DT context, recalling the DT-related experiences as manipulations or participating in DT itself is less likely to trigger one's moral licensing processes. Rather, doing so will generate self-consistency effects in this context. Perhaps, to evoke moral licensing processes in this context, researchers may need to devise other ways that can offer moral self-validation for participants. For instance, participants who receive a documented certificate after the completion of DT might use such a tangible certificate to self-validate themselves. Alternatively, as the prior study found the moral licensing effects from recalling positive experiences with minority group members (Bradley-Geist et al., 2012), future research may focus on whether participants can morally license their recent interactions with minority group members, which might influence DT backlash. Creating a further understanding of moral licensing processes in DT backlash will provide a meaningful step toward a complete picture of the phenomenon.

There was some evidence that moral licensing processes occurred for research participants. Although it only showed weak effect sizes on psychological entitlement and willingness to help minority group members in Study 1, it still shows that moral licensing can occur. This tendency occurred when people who prefer social hierarchy recalled the number of DT attended or the positiveness of DEI policies, and also when people who believe the procedure during DT was fair recalled the number of DT attended. Such results suggest that the self-consistency and moral licensing effects of recalling could co-occur to influence different DT backlash manifestations, meaning that the two effects may be elicited in different manifestations at the same time. For example, one might experience moral licensing effects on the behavioral manifestation of DT backlash, while the same person may reduce explicit prejudice as a result of recalling. This suggests that having mixed effects from recalling DT-related experiences is plausible, even though my finding also suggests that self-consistency effects still may be more prevalent from recalling the DT-related experiences.

To further identify boundary conditions that can tease out when self-consistency would occur and when moral licensing would take place after recalling DT-related experiences, future research may focus on training characteristics, such as whether the recalled DT was mandatory to participate. The effectiveness of mandatory or voluntary DT has been an important research question in the literature (Bezrukova et al., 2016; Kulik & Roberson, 2008), and the meta-analysis found that the mandatory training did not significantly differ from voluntary participation. However, moral licensing theory implies that people may be more likely to license their voluntary training participation rather than mandatory training experiences because the former has higher moral values. Thus, when asked to recall voluntary DT experiences, people may be more likely to show DT backlash based on moral licensing than participants who recalled mandatory DT experiences. On the other hand, recalling voluntary DT experiences may also equally motivate individuals to engage more with the subsequent DT. Examining recalling past DT experiences affects licensing processes or self-consistency processes in which specific DT backlash manifestations can provide unique and important insights for the important research question.

Relatedly, the testing and results of the role of justice perceptions and personality traits can also guide future research directions for the moral licensing theory. Perhaps due to the moral licensing theory's origin in social psychology, the moral licensing theory literature tends to focus more on contexts and situations that elicit moral licensing, leaving the role of individual differences unquestioned. By incorporating individual differences in my hypotheses, I demonstrated how one's stable features interacted differently with the manipulations. Despite the lack of support provided for the moral licensing hypotheses that were tested, the study's findings raise important questions that should be addressed in future research. The results of the self-consistency effects suggest that personality traits included in this study might have been too biased against DT and minority group members that individuals with such traits do not even consider DT experiences as moral or ethical, thwarting moral licensing effects. Thus, future research may test how personality traits or individual differences that are not directly related to prejudice affect one's likelihood to engage in moral licensing. For example, one's goal orientations which measure whether a person is learning-oriented or performance-oriented could moderate the relations between manipulations and DT backlash. This is because a person with performance orientation may attempt to demonstrate their competence and avoid looking bad (Dweck, 1988) by relying on past DT experiences, and such tendencies may be more likely to elicit moral licensing processes.

In addition, individuals with high unfairness perceptions, high SDO, and high BJW demonstrated reduced DT backlash manifestations after recalling their past good deeds. This result indicates that people with personality traits that are negative and directly related to prejudice may not engage in moral licensing processes. I conjecture that personality traits, which are more positive and self-focused (e.g., high levels in core self-evaluations) might provide better predictions for who might be more frequently experiencing moral licensing processes because people with those traits may possess readily available information about self that could be easier to self-license.

Limitations

Although the studies that I conducted yielded some interesting and potentially important findings, it is not without limitations. First, the manipulations did not generate intended moral licensing processes to the extent that I anticipated. A possible explanation is that participants did not consider DEI values or DT to be a moral issue. The manipulations used in both studies presumed that participants would view DT or DEI issues as involving moral or ethical values. However, there is a possibility that participants may not always perceive DEI issues or DT as moral issues. If participants did not view DT as moral, then recalling their previous experiences of DT or the DEI policies at the current workplace may not produce moral licensing effects because those experiences may not provide self-validations for participants. This might be a reason why the moral licensing processes were less emphasized in both studies. In future studies, researchers should include manipulation checks that test whether the manipulations are tapping their moral boundary or change manipulations to involve stronger fidelity to arouse one’s moral identity regarding DEI values. For example, researchers can first ask whether they view DEI issues and DT as moral issues to ensure the possibility that participants may use their good deeds regarding DEI issues as moral credentials. Alternatively, researchers may present vignettes or scenarios that highlight moral dilemmas in DEI issues and ask participants to make decisions on such issues. If such decisions were moral, then the participants might be more likely to license their moral decisions as in Monin and Miller (2001)’s study.

To measure the behavioral manifestations of DT backlash, I asked participants to rate the extent to which they would help or hire minority group members. Because these measures are only behavioral tendencies rather than behaviors, an additional limitation is the possibility that respondents may behave differently in reality from how they responded to the questionnaires. Furthermore, due to the nature of the experimental setting, participants read a scenario about a hiring situation and then self-reported their willingness to hire minority group members. Reading a vignette may not be realistic for participants. To aid this issue, researchers may assess the behavioral manifestations of DT backlash by asking their peers or coworkers to report the focal trainees' behavioral manifestations of DT backlash such as observed focal employees' helping behavior and social undermining behavior. Also, to provide respondents with more realistic hiring situations, future studies may extend a fictitious hiring context where respondents have to review fictitious resumes and select one of the candidates from the pool of majority and minority group members with equivalent competencies.

Among the various manifestations of DT backlash, the manipulations did not significantly affect implicit prejudice. These results may be partly due to the fact that the IAT used in the present studies only tests participants' implicit association regarding two specific races (e.g., European American and African American) even though DEI values and the DT video that participants watched tend to provide much broader categories. Also, I measured explicit prejudice and willingness to hire minority group members only regarding African Americans. As a result, the present studies did not examine the full spectrum of one's implicit and explicit prejudice, and hiring decisions including various social categories such as gender, age, sexual orientation, and disability. Including different categories in IAT and explicit prejudice should provide more detailed aspects about DT backlash in future studies. Also, to provide a complete picture of and more rigorous testing of DT backlash, future studies should include measures that match the breadth of DEI values utilized in the study.

Lastly, except for the pretest and posttest of IAT, I did not measure the pretests of dependent variables. As a result, the results of the present study can explain the differences created by the manipulations across groups, but it cannot be concluded that the manipulations created significant within-person differences. Also, because interpreting the results of the moral licensing processes often rely on the sequences of behavior, it is often difficult for researchers to conclude whether moral licensing processes have occurred without knowing the baseline results (i.e., pretest results, Mullen & Miller, 2016). If future studies could include the pretests of DT backlash variables, doing so will increase the internal validity of the experiments and find a more rigorous indication of DT backlash manifestations.

Practical Implications

In addition to the theoretical implications, this research also offers important guidance for practitioners. First, in contrast to studies that found moral licensing effects based on recalling previous good behavior (e.g., Blanken et al., 2012), this dissertation found that recalling various aspects of DT, including past experiences, the number of DT taken, and the positive aspects of DEI policies, can weaken DT backlash. Such results suggest that managers and organizations can play a vital role in communicating and linking trainees' previous DT experiences to the subsequent DT. Managers and organizations can also focus on refreshing what trainees learned in the last DT, how DT fits with overall DEI initiatives that the organization is pursuing, and how DEI policies provided by the organization fit with DT, and such strategies should motivate trainees before participating in DT. More importantly, these actions will motivate trainees with high prejudice and unfairness perceptions, who could be the primary intended targets of DT.

Thus, to achieve DT's goal of improving DT effectiveness by reducing DT backlash, it is imperative to prompt trainees about their past DT experiences and how such experiences can improve their learning after DT and reduce prejudice against minority group members.

Second, when evaluating DT effectiveness or training outcomes, applying the definition and measurements of DT backlash can be highly practical for organizations and managers. Because DT backlash was defined as it can independently occur regardless of DT outcomes, knowing and ensuring whether such backfiring effects were generated or not could be valuable information in evaluating DT. For example, if one type of DT is effective in terms of increasing learning outcomes and awareness, but if it also generates DT backlash, then managers may not choose to use it next time. Managers may incorporate the measurements or questionnaires that I used in this study, such as explicit prejudice and IATs, to identify whether a particular manifestation of DT backlash occurred. If none of the DT backlash measures were incorporated in evaluating DT, the practitioners should interpret the outcomes of DT carefully because the backlash has occurred. Also, practitioners should be cautious about interpreting the negative reactions measured at the immediate end of training because this may not just indicate DT was ineffective but also that DT backlash occurred. Thus, by using my definition and measures, the organization may be better able to evaluate whether DT was effective or not.

Lastly, although my dissertation started out to find the psychological mechanism that can increase DT backlash, I found that recalling DT related experiences can significantly reduce DT backlash. I also found that taking DT, even a very short version, can increase moral credentials and metacognitive activity. These results highly suggest that although researchers often argue that DT is often a driver that increases prejudice and is largely ineffective (Dobbin & Kalev, 2006, 2016, 2018), DT can be effective in generating positive outcomes. Such outcomes were found to be greatly enhanced if trainees could link their past learning to the current DT. More importantly, such recalling was more beneficial for trainees with deep-seated biased and feeling unfair about DEI values. Future studies may focus on whether such effects can influence one's DT transfer outcomes, such as reduced DT backlash transfer to the daily work domain. In conclusion, this study highlights that if DT is managed in a way that redirects trainees' attention and awareness toward DEI values, it can enhance its outcomes by reducing DT backlash.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*(1), 3–27. https://doi.org/10.1037/0033-2909.102.1.3

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267–299). Academic Press. https://doi.org/10.1016/S0065-2601(08)60108-2

Ahmad, M. G., Klotz, A. C., & Bolino, M. C. (2021). Can good followers create unethical leaders? How follower citizenship leads to leader moral licensing and unethical behavior. *Journal of Applied Psychology*, *106*(9), 1374–1390. https://doi.org/10.1037/apl0000839

Alderfer, C. P. (1992). Changing race relations embedded in organizations: Report on a long-term project with the XYZ Corporation. In S. E. Jackson (Ed.), *Diversity in the workplace: Human resources initiatives* (pp. 138–166). Guilford Press.

Alderfer, C. P., Alderfer, C. J., Bell, E. L., & Jones, J. (1992). The race relations competence workshop: Theory and results. *Human Relations*, *45*(12), 1259–1291. https://doi.org/10.1177/001872679204501202

Allport, G. W. (1929). The composition of political attitudes. *American Journal of Sociology*, *35*(2), 220–238. https://doi.org/10.1086/214980

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Clark University Press.

Alliger, G. M., Tannenbaum, S. I., Bennett Jr, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, *50*(2), 341–358. https://doi.org/10.1111/j.1744-6570.1997.tb00911.x

Alter, K. J., & Zürn, M. (2020). Conceptualising backlash politics: Introduction to a special issue on backlash politics in comparison. *British Journal of Politics and International Relations*, *22*(4), 563–584. https://doi.org/10.1177/1369148120947958

Anand, R., & Winters, M. F. (2008). A retrospective view of corporate diversity training from 1964 to the present. *Academy of Management Learning & Education*, *7*(3), 356–372. https://doi.org/10.5465/amle.2008.34251673

Ancis, J. R., & Szymanski, D. M. (2001). Awareness of White privilege among White counseling trainees. *The Counseling Psychologist*, *29*(4), 548–569. https://doi.org/10.1177/0011000001294005

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. https://doi.org/10.1037/0033-295X.89.4.369

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist: Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology*, *95*, 918–932. https://doi.org/10.1037/a0011990

Aquino, K., & Reed, A. II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267. https://doi.org/10.1037/0033-2909.115.2.243

Beaver, W. (1995). Let's stop diversity training and start managing for diversity. *Industrial Management*, *37*(4), 7–9.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: What we know and where we should go. *Journal of Applied Psychology*, *102*(3), 305–323. https://doi.org/10.1037/apl0000142

Bergman, M. E., & Salter, P. (2013). Backlash! What it is, where it comes from, and how we can fix it. *Industrial and Organizational Psychology*, *6*(4), 442–450. https://doi.org/10.1111/iops.12082

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*(2), 410–424. https://doi.org/10.1037/0021-9010.92.2.410

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin*, *142*(11), 1227–1274. https://doi.org/10.1037/bul0000067

Bezrukova, K., Jehn, K. A., & Spell, C. S. (2012). Reviewing diversity training: Where we have been and where we should go. *Academy of Management Learning & Education*, *11*(2), 207–227. https://doi.org/10.1037/a0023684

Blanken, I., Van De Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, *41*(4), 540–558. https://doi.org/10.1177/0146167215572134

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*(3), 567–582. https://doi.org/10.1037/a0014665

Bobocel, D. R., Son Hing, L. S., Davey, L. M., Stanley, D. J., & Zanna, M. P. (1998). Justice-based opposition to social policies: Is it genuine? *Journal of Personality and Social Psychology*, *75*(3), 653–669. https://doi.org/10.1037/0022-3514.75.3.653

Bohnet, I. (2016). *What works*. Harvard university press.

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., & McKenna, C. (2010). Moral credentialing by association: The importance of choice and relationship closeness. *Personality and Social Psychology Bulletin*, *36*(11), 1564–1575. https://doi.org/10.1177/0146167210385920

Brannon, T. N., Carter, E. R., Murdock-Perriera, L. A., & Higginbotham, G. D. (2018). From backlash to inclusion for all: Instituting diversity efforts to maximize benefits across group lines. *Social Issues and Policy Review*, *12*(1), 57–90.

Branscombe, N. R., & Wann, D. L. (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology*, *24*(6), 641–657. https://doi.org/10.1002/ejsp.2420240603

Brehm, J. W. (1966). *A theory of psychological reactance.* Academic Press.

Brehm, S. S. & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control.* Academic Press.

Brewis, D. N. (2019). Duality and fallibility in practices of the self: The ‘inclusive subject’ in diversity training. *Organization Studies*, *40*(1), 93–114. https://doi.org/10.1177/0170840618765554

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell& E. M. Markman (Eds.), *Handbook of Child psychology: Vol. 3. Cognitive development* (4th ed., pp. 77-166). New York: Wiley.

Burke, R.J., Black, S. (1997). Save the males: Backlash in organizations. In Burke, R. J. (eds.) *Women in Corporate Management* (pp. 61–70). Springer. https://doi.org/10.1007/978-94-011-5610-3_7

Bushman, B. J., & Stack, A. D. (1996). Forbidden fruit versus tainted fruit: Effects of warning labels on attraction to television violence. *Journal of Experimental Psychology: Applied*, *2*(3), 207–226. https://doi.org/10.1037/1076-898X.2.3.207

Campbell, D. T. (1965). Ethnocentric and other altruistic motives. In D. Levine (Ed.), *Nebraska Symposium on motivation* (pp. 283–311). University of Nebraska Press.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Chung, S., Zhan, Y., Noe, R. A., & Jiang, K. (2022). Is it time to update and expand training motivation theory? A meta-analytic review of training motivation research in the 21st century. *Journal of Applied Psychology*, *107*(7), 1150–1179. https://doi.org/10.1037/apl0000901

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*(3), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*(3), 425–445. https://doi.org/10.1037/0021-9010.86.3.425

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin*, *38*(7), 907–919. https://doi.org/10.1177/0146167212442394

Correll, J., Park, B., & Smith, J. A. (2008). Situations colorblind and multicultural prejudice reduction strategies in high-conflict. *Group Processes & Intergroup Relations*, *11*, 471–491. https://doi.org/10.1177/1368430208095401

Crandall, C. S., & Cohen, C. (1994). The personality of the stigmatizer: Cultural world view, conventionalism, and self-esteem. *Journal of Research in Personality*, *28*(4), 461–480. https://doi.org/10.1006/jrpe.1994.1033

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Chavez, C.I. and Weisinger, J.Y. (2008), Beyond diversity training: A social infusion for cultural inclusion. *Human Resource Management*, *47*, 331–350. https://doi.org/10.1002/hrm.20215

D'Andrea, M., Daniels, J. & Heck, R. (1991). Evaluating the impact of multicultural counseling training. *Journal of Counseling & Development, 70*(1), 143–150. https://doi.org/10.1002/j.1556-6676.1991.tb01576.x

Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86*(5), 1022–1033. https://doi.org/10.1037/0021-9010.86.5.1022

Devine, P. G., & Ash, T. L. (2022). Diversity training goals, limitations, and promise: a review of the multidisciplinary literature. *Annual Review of Psychology*, *73*(1), 403–429. https://doi.org/10.1146/annurev-psych-060221-122215

Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, *72*(2), 144–168. https://doi.org/10.1080/03637750500111815

Dobbin, F., & Kalev, A. (2018). Why doesn't diversity training work? The challenge for industry and academia. *Anthropology Now*, *10*(2), 48–55. https://doi.org/10.1080/19428200.2018.1493182

Dobbin, F., Schrage, D., & Kalev, A. (2015). Rage against the iron cage: The varied effects of bureaucratic personnel reforms on diversity. *American Sociological Review*, *80*(5), 1014–1044. https://doi.org/10.1177/0003122415596416

Doosje, B., Ellemers, N., & Spears, R. (1999). Commitment and intergroup behavior. In N. Ellemers, R. Spears, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 84–106). Blackwell Science.

Doosje, B., Spears, R., & Ellemers, N. (2002). The dynamic and determining role of ingroup identification: Responses to anticipated and actual changes in the intergroup status hierarchy. *British Journal of Social Psychology, 41*(1), 57–76.

Dover, T. L., Kaiser, C. R., & Major, B. (2020). Mixed signals: The unintended effects of diversity initiatives. *Social Issues and Policy Review*, *14*(1), 152–181. https://doi.org/10.1111/sipr.12059

Downey, S. N., van der Werff, L., Thomas, K. M., & Plaut, V. C. (2015). The role of diversity practices and inclusion in promoting trust and employee engagement. *Journal of Applied Social Psychology*, *45*(1), 35–44. https://doi.org/10.1111/jasp.12273

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *Journal of Applied Psychology, 100*(2), 343–359. https://doi.org/10.1037/a0037908

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring whites. *Journal of Experimental Social Psychology*, *45*(3), 590–593. https://doi.org/10.1016/j.jesp.2009.02.001

Effron, D. A., & Conway, P. (2015). When virtue leads to villainy: Advances in research on moral self-licensing. *Current Opinion in Psychology*, *6*, 32–35. https://doi.org/10.1016/j.copsyc.2015.03.017

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions?. *Personality and Social Psychology Bulletin*, *36*(12), 1618–1634. https://doi.org/10.1177/0146167210385922

Egan, M. L., & Bendick Jr, M. (2008). Combining multicultural management and diversity into one course on cultural competence. *Academy of Management Learning & Education*, *7*(3), 387–393. https://doi.org/10.5465/amle.2008.34251675

Eibach, R. P., & Keegan, T. (2006). Free at last? Social dominance, loss aversion, and white and black Americans' differing assessments of racial progress. *Journal of Personality and Social Psychology, 90*(3), 453–467. https://doi.org/10.1037/0022-3514.90.3.453

Feather, N. T. (1984). Protestant Ethic, conservatism, and values. *Journal of Personality and Social Psychology, 46*(5), 1132–1141. https://doi.org/10.1037/0022-3514.46.5.1132

Fishbach, A., Zhang, Y., & Koo, M. (2009). The dynamics of self-regulation. *European Review of Social Psychology, 20*(1), 315–344. https://doi.org/10.1080/10463280903275375

Fishbach, A., Koo, M., & Finkelstein, S. R. (2014). Motivation resulting from completed and missing actions. In A. Olson & M. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 50, pp. 257–307). Academic Press. https://doi.org/10.1016/B978-0-12-800284-1.00005-9

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology*, *83*(2), 218–233. https://doi.org/10.1037/0021-9010.83.2.218

Fox, S., & Spector, P. E. (1999). A model of work frustration–aggression. *Journal of Organizational Behavior*, *20*(6), 915–931. https://doi.org/10.1002/(SICI)1099-1379(199911)20:6<915::AID-JOB918>3.0.CO;2-6

Furnham, A., & Gunter, B. (1984). Just world beliefs and attitudes towards the poor. *British Journal of Social Psychology*, *23*(3), 265–269. https://doi.org/10.1111/j.2044-8309.1984.tb00637.x

Furnham, A. (2003). Belief in a just world: Research progress over the past decade. *Personality and Individual Differences, 34*(5), 795–817. https://doi.org/10.1016/S0191-8869(02)00072-7

Furnham, A., & Procter, E. (1989). Belief in a just world: Review and critique of the individual difference literature. *British Journal of Social Psychology*, *28*(4), 365–384. https://doi.org/10.1111/j.2044-8309.1989.tb00880.x

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. https://doi.org/10.1037/0022-3514.78.4.708

Garcia-Marques, L., Santos, A. S. C., & Mackie, D. M. (2006). Stereotypes: Static Abstractions or Dynamic Knowledge Structures? *Journal of Personality and Social Psychology, 91*(5), 814–831. https://doi.org/10.1037/0022-3514.91.5.814

Giangreco, A., Carugati, A., Sebastiano, A., & Della Bella, D. (2010). Trainees' reactions to training: shaping groups and courses for happier trainees. *International Journal of Human Resource Management*, *21*(13), 2468–2487. https://doi.org/10.1080/09585192.2010.516598

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., & Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science, 58*(1), 179–187. https://doi.org/10.1073/pnas.112089310

Graupmann, V., Jonas, E., Meier, E., Hawelka, S., & Aichhorn, M. (2012). Reactance, the self, and its group: When threats to freedom come from the ingroup versus the outgroup. *European Journal of Social Psychology*, *42*(2), 164–173. https://doi.org/10.1002/ejsp.857

Green, E. (2000, November 3). Most colleges require diversity education, survey finds. *The Chronicle of Higher Education*. https://www.chronicle.com/article/most-colleges-require-diversity-education-survey-finds/

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Gündemir, S., Homan, A. C., Usova, A., & Galinsky, A. D. (2017). Multicultural meritocracy: The synergistic benefits of valuing diversity and merit. *Journal of Experimental Social Psychology, 73*, 34–41. https://doi.org/10.1016/j.jesp.2017.06.002

Hanover, J. M., & Cellar, D. F. (1998). Environmental factors and the effectiveness of workforce diversity training. *Human Resource Development Quarterly*, *9*(2), 105–124. https://doi.org/10.1002/hrdq.3920090203

Heilman, M. (1994). Affirmative action: Some unintended consequences for working women. In *Research in organizational behavior* (pp. 125–169). JAI Press.

Heilman, M. E., & Welle, B. (2006). Disadvantaged by diversity? The effects of diversity goals on competence perceptions. *Journal of Applied Social Psychology*, *36*(5), 1291–1319. https://doi.org/10.1111/j.0021-9029.2006.00043.x

Hershcovis, M. S., Turner, N., Barling, J., Arnold, K. A., Dupré, K. E., Inness, M., LeBlanc, M. M., & Sivanathan, N. (2007). Predicting workplace aggression: A meta-analysis. *Journal of Applied Psychology*, *92*(1), 228–238. https://doi.org/10.1037/0021-9010.92.1.228

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Hogg, M. A., & Terry, D. J. (2000). The dynamic, diverse, and variable faces of organizational identity. *Academy of Management Review*, *25*(1), 150–152. https://doi.org/10.5465/amr.2000.27711645

Holladay, C. L., Knight, J. L., Paige, D. L., & Quiñones, M. A. (2003). The influence of framing on attitudes toward diversity training. *Human Resource Development Quarterly*, *14*(3), 245–263. https://doi.org/10.1002/hrdq.1065

Holladay, C. L., & Quiñones, M. A. (2005). Reactions to diversity training: An international comparison. *Human Resource Development Quarterly*, *16*(4), 529–545. https://doi.org/10.1002/hrdq.1154

Holladay, C. L., & Quiñones, M. A. (2008). The influence of training focus and trainer characteristics on diversity training effectiveness. *Academy of Management Learning & Education*, *7*(3), 343–354. https://doi.org/10.5465/amle.2008.34251672

Hong, S.-M., & Faedda, S. (1996). Refinement of the Hong psychological reactance scale. *Educational and Psychological Measurement*, *56*(1), 173–182. https://doi.org/10.1177/0013164496056001014

Hong, S.-M., & Page, S. (1989). A psychological reactance scale: Development, factor structure and reliability. *Psychological Reports*, *64*(3), 1323–1326. https://doi.org/10.2466/pr0.1989.64.3c.1323

Hood, J. N., Muller, H. J., & Seitz, P. (2001). Attitudes of Hispanics and Anglos surrounding a workforce diversity intervention. *Hispanic Journal of Behavioral Sciences*, *23*(4), 444–458.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Jackson, L. C. (1999). Ethnocultural resistance to multicultural training: Students and faculty. *Cultural Diversity and Ethnic Minority Psychology*, *5*(1), 27–36. https://doi.org/10.1037/1099-9809.5.1.27

Jones, K. P., Peddie, C. I., Gilrane, V. L., King, E. B., & Gray, A. L. (2016). Not so subtle: A meta-analytic investigation of the correlates of subtle and overt discrimination. *Journal of Management*, *42*(6), 1588–1613. https://doi.org/10.1177/0149206313506466

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of Personality and Social Psychology*, *104*(3), 504–519. https://doi.org/10.1037/a0030838

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*(4), 589–617. https://doi.org/10.1177/000312240607100404

Kaplan, D. M. (2006). Can diversity training discriminate? Backlash to lesbian, gay, and bisexual diversity initiatives. *Employee Responsibilities and Rights Journal*, *18*(1), 61–72. https://doi.org/10.1007/s10672-005-9005-4

Karp, H. B., & Sammour, H. Y. (2000). Workforce diversity: Choices in diversity training programs & dealing with resistance to diversity. *College Student Journal*, *34*(3), 451–451.

Kernahan, C., & Davis, T. (2007). Changing perspective: How learning about racism influences student awareness and emotion. *Teaching of Psychology*, *34*(1), 49–52. https://doi.org/10.1080/00986280709336651

Khan, U., & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research*, *43*(2), 259–266. https://doi.org/10.1509/jmkr.43.2.259

Kidder, D. L., Lankau, M. J., Chrobot-Mason, D., Mollica, K. A., & Friedman, R. A. (2004). Backlash toward diversity initiatives: Examining the impact of diversity program justification, personal and group outcomes. *International Journal of Conflict Management*, *15*(1), 77–102. https://doi.org/10.1108/eb022908

Kim, S. Y., Levine, T., & Allen, M. (2013). Comparing separate process and intertwined models for reactance. *Communication Studies*, *64*(3), 273–295. https://doi.org/10.1080/10510974.2012.755639

Kirkpatrick, D. L. (1994). *Evaluating training programs: the four levels*. Berrett-Koehler.

Kitayama, S., Markus, H. R., & Kurokawa, M. (2000). Culture, emotion, and well-being: Good feelings in Japan and the United States. *Cognition & Emotion*, *14*(1), 93–124. https://doi.org/10.1080/026999300379003

Kitayama, S., Mesquita, B., & Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology*, *91*(5), 890–903. https://doi.org/10.1037/0022-3514.91.5.890

Klotz, A. C., & Bolino, M. C. (2013). Citizenship and counterproductive work behavior: A moral licensing view. *Academy of Management Review*, *38*(2), 292–306. https://doi.org/10.5465/amr.2011.0109

Koo, M., & Fishbach, A. (2010). Climbing the goal ladder: How upcoming actions increase level of aspiration. *Journal of Personality and Social Psychology*, *99*(1), 1–13. https://doi.org/10.1037/a0019443

Kouchaki, M. (2011). Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of Personality and Social Psychology*, *101*(4), 702–715. https://doi.org/10.1037/a0024552

Kowal, E., Franklin, H., & Paradies, Y. (2013). Reflexive antiracism: A novel approach to diversity training. *Ethnicities*, *13*(3), 316–337. https://doi.org/10.1177/1468796812472885

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311–328. https://doi.org/10.1037/0021-9010.78.2.311

Kulik, C. T., Perry, E. L., & Bourhis, A. C. (2000). Ironic evaluation processes: Effects of thought suppression on evaluations of older job applicants. *Journal of Organizational Behavior*, *21*(6), 689–711. https://doi.org/10.1002/1099-1379(200009)21:6<689::AID-JOB52>3.0.CO;2-W

Kulik, C. T., & Roberson, L. (2008). Common goals and golden opportunities: Evaluations of diversity education in academic and organizational settings. *Academy of Management Learning & Education*, *7*(3), 309–331. https://doi.org/10.5465/amle.2008.34251670

Lacerenza, C. N., Reyes, D. L., Marlow, S. L., Joseph, D. L., & Salas, E. (2017). Leadership training design, delivery, and implementation: A meta-analysis. *Journal of Applied Psychology, 102*(12), 1686–1718. https://doi.org/10.1037/apl0000241

Law, D. Y. (1998). *An evaluation of a cultural diversity training program*. Auburn University.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, *22*(12), 1472–1477. https://doi.org/10.1177/0956797611427918

Lerner, M. J. (1965). Evaluation of performance as a function of performer's reward and attractiveness. *Journal of Personality and Social Psychology, 1*(4), 355–360. https://doi.org/10.1037/h0021806

Lerner M. J. (1980). *The belief in a just world: A fundamental delusion*. Plenum.

Lerner, M. J., & Miller, D. T. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin, 85*(5), 1030–1051. https://doi.org/10.1037/0033-2909.85.5.1030

Leslie, L. M. (2019). Diversity initiative effectiveness: A typological theory of unintended consequences. *Academy of Management Review*, *44*(3), 538–563. https://doi.org/10.5465/amr.2017.0087

Leslie, L. M., Mayer, D. M., & Kravitz, D. A. (2014). The stigma of affirmative action: A stereotyping-based theory and meta-analytic test of the consequences for performance. *Academy of Management Journal*, *57*(4), 964–989. https://doi.org/10.5465/amj.2011.0940

LeVine, R. A., & Campbell, D. T. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior*. John Wiley & Sons.

Levinthal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In K. S. Gergen, M. S. Green berg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). Plenum.

Levinthal, G. S., Karuza, J., & Fry, W. R. (1980). Beyond fairness: A theory of allocation preferences. In G. Mikula (Ed.), *Justice and social interaction* (pp. 167–218). Springer.

Liang, L. H., Coulombe, C., Skyvington, S., Brown, D. J., Ferris, D. L., & Lian, H. (2022). License to Retaliate: Good Deeds as a Moral License for Misdeeds in Reaction to Abusive Supervision. *Human Performance, 35*(2), 94–112. https://doi.org/10.1080/08959285.2022.2032069

Liberman, N., & Förster, J. (2000). Expression after suppression: A motivational explanation of postsuppressional rebound. *Journal of Personality and Social Psychology, 79*(2), 190–203. https://doi.org/10.1037/0022-3514.79.2.190

Lin, S.-H. (J.), Ma, J., & Johnson, R. E. (2016). When ethical leader behavior breaks bad: How ethical leader behavior can turn abusive via ego depletion and moral licensing. *Journal of Applied Psychology, 101*(6), 815–830. https://doi.org/10.1037/apl0000098

Lindsay, C. (1994). Things that go wrong in diversity training: Conceptualization and change with ethnic identity models. *Journal of Organizational Change Management, 7*(6), 18–33. https://doi.org/10.1108/09534819410072683

Lindsey, A., King, E., & Amber, B. (2020). Diversity Training Effectiveness: Affective mechanisms, motivational drivers, individual difference moderators, and contextual boundary conditions. In D. L. Stone, J. H. Dulebohn, & K. M. Lukaszewski (Eds.), *Diversity and inclusion in organizations* (pp. 137–164). Information Age.

Lipkus, I. (1991). The construction and preliminary validation of a global belief in a just world scale and the exploratory analysis of the multidimensional belief in a just world scale. *Personality and Individual differences*, *12*(11), 1171–1178. https://doi.org/10.1016/0191-8869(91)90081-L

Lipman, J. (2018, January 25). How diversity training infuriates men and fails women. *Time*. https://time.com/5118035/diversity-training-infuriates-men-fails-women/

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology, 67*(5), 808–817. https://doi.org/10.1037/0022-3514.67.5.808

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of personality and social psychology*, *79*(4), 602–616. https://doi.org/10.1037/0022-3514.79.4.602

Majer, J. M., Trötschel, R., Galinsky, A. D., & Loschelder, D. D. (2020). Open to offers, but resisting requests: How the framing of anchors affects motivation and negotiated outcomes. *Journal of Personality and Social Psychology*, *119*(3), 582–599. https://doi.org/10.1037/pspi0000210

Mann, N. H., & Kawakami, K. (2012). The long, steep path to equality: Progressing on egalitarian goals. *Journal of Experimental Psychology: General*, *141*(1), 187–197. https://doi.org/10.1037/a0025602

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people? *Psychological Science*, *21*(4), 494–498. https://doi.org/10.1177/0956797610363538

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). Academic Press.

McGregor, J. (2016, July 1). To improve diversity, don't make people go to diversity training. Really. *Washington Post*. https://www.washingtonpost.com/news/on-leadership/wp/2016/07/01/to-improve-diversity-dont-make-people-go-to-diversity-training-really-2/

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344–357. https://doi.org/10.1111/j.1751-9004.2010.00263.x

Miller, D. A., Smith, E. R., & Mackie, D. M. (2004). Effects of intergroup contact and political predispositions on prejudice: Role of intergroup emotions. *Group Processes & Intergroup Relations*, *7*(3), 221–237. https://doi.org/10.1177/1368430204046109

Miller, D. T., & Effron, D. A. (2010). Psychological license: When it is needed and how it functions. In P. Z. Mark, & M. O. James (Eds.), *Advances in experimental social psychology* (pp. 115–155). Academic Press. https://doi.org/10.1016/S0065-2601(10)43003-8

Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., & Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, *33*(2), 219–240. https://doi.org/10.1111/j.1468-2958.2007.00297.x

Mobley, M., & Payne, T. (1992). Backlash! The challenge to diversity training. *Training & Development*, *46*(12), 45–52.

Mollica, K. A. (2003). The influence of diversity context on white men's and racial minorities' reactions to disproportionate group harm. *The Journal of Social Psychology*, *143*(4), 415–431. https://doi.org/10.1080/00224540309598454

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33–43. https://doi.org/10.1037/0022-3514.81.1.33

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, *67*(1), 363–385. https://doi.org/10.1146/annurev-psych-010213-115120

Nelson, T. E., Acker, M., & Manis, M. (1996). Irrepressible stereotypes. *Journal of Experimental Social Psychology*, *32*(1), 13–38. https://doi.org/10.1006/jesp.1996.0002

Neville, H., & Furlong, M. (1994). The impact of participation in a cultural awareness program on the racial attitudes and social behaviors of first-year college students. *Journal of College Student Development*, *35*(5), 371–377.

Neville, H. A., Awad, G. H., Brooks, J. E., Flores, M. P., & Bluemel, J. (2013). Color-blind racial ideology: Theory, training, and measurement implications in psychology. *American Psychologist*, *68*(6), 455–466. https://doi.org/10.1037/a0033282

Nisan, M. (1991). The moral balance model: Theory and research extending our understanding of moral choice and deviation. In W. M. Kurtines & J. L. Gerwitz (Eds.), *Handbook of moral behavior and development*, (pp. 213–249). Lawrence Erlbaum.

Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review*, *11*(4), 736–749. https://doi.org/10.2307/258393

Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39*(3), 497–523. https://doi.org/10.1111/j.1744-6570.1986.tb00950.x

Noe, R. A., & Wilk, S. L. (1993). Investigation of the factors that influence employees' participation in development activities. *Journal of Applied Psychology*, *78*(2), 291–302. https://doi.org/10.1037/0021-9010.78.2.291

Norton, M. I., & Sommers, S. R. (2011). Whites see racism as a zero-sum game that they are now losing. *Perspectives on Psychological science*, *6*(3), 215–218. https://doi.org/10.1177/1745691611406922

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction. *Psychological Science*, *17*, 949 –953. https://doi.org/10.1111/j.1467-9280.2006.01810.x

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to action research. *Journal of Social Issues*, *62*(3), 577–595. https://doi.org/10.1111/j.1540-4560.2006.00474.x

Pendry, L. F., Driscoll, D. M., & Field, S. C. (2007). Diversity training: Putting theory into practice. *Journal of Occupational and Organizational Psychology*, *80*(1), 27–50. https://doi.org/10.1348/096317906X118397

Pietri, E. S., Hennes, E. P., Dovidio, J. F., Brescoll, V. L., Bailey, A. H., Moss-Racusin, C. A., & Handelsman, J. (2019). Addressing unintended consequences of gender diversity interventions on women's sense of belonging in STEM. *Sex Roles*, *80*(9), 527–547.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/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*(6), 486–501. https://doi.org/10.1006/jesp.2001.1478

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Pratto, F., Sidanius, J., & Levin, S. (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European Review of Social Psychology*, *17*(1), 271–320. https://doi.org/10.1080/10463280601055772

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., & Ahmad, M. (1987). Generation and precision of elaboration: Effects on intentional and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 291–300. https://doi.org/10.1037/0278-7393.13.2.291

Probst, T. M. (2003). Changing attitudes over time: Assessing the effectiveness of a workplace diversity course. *Teaching of Psychology*, *30*(3), 236–239. https://doi.org/10.1207/S15328023TOP3003_09

Quick, B. L., & Stephenson, M. T. (2008). Examining the role of trait reactance and sensation seeking on perceived threat, state reactance, and reactance restoration. *Human Communication Research*, *34*(3), 448–476. https://doi.org/10.1111/j.1468-2958.2008.00328.x

Quiñones, M. A. (1995). Pretraining context effects: Training assignment as feedback. *Journal of Applied Psychology*, *80*(2), 226–238. https://doi.org/10.1037/0021-9010.80.2.226

Ragins, B. R., & Ehrhardt, K. (2021). Gaining perspective: The impact of close cross-race friendships on diversity training and education. *Journal of Applied Psychology*, *106*(6), 856–881. https://doi.org/10.1037/apl0000807

Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human Communication Research*, *39*(1), 47–73. https://doi.org/10.1111/j.1468-2958.2012.01443.x

Ratner, R. K., & Miller, D. T. (2001). The norm of self-interest and its effects on social action. *Journal of Personality and Social Psychology*, *81*(1), 5–16. https://doi.org/10.1037/0022-3514.81.1.5

Royal Bank of Canada. (2020, August 6). When More People Speak Up, More People Listen [Video]. YouTube. https://www.youtube.com/watch?v=b1nJqpqgzR0&t=1s

Richard, O. C., Barnett, T., Dwyer, S., & Chadwick, K. (2004). Cultural diversity in management, firm performance, and the moderating role of entrepreneurial orientation dimensions. *Academy of Management Journal*, *47*(2), 255–266. https://doi-org.proxy2.cl.msu.edu/10.5465/20159576

Rim, Y. (1988). Attitudes and the confluence model. *Small Group Behavior*, *19*(1), 153–161.

Roberson, L., Kulik, C. T., & Pepper, M. B. (2001). Designing effective diversity training: Influence of group composition and trainee experience. *Journal of Organizational Behavior*, *22*(8), 871–885. https://doi.org/10.1002/job.117

Roberson, Q., Ryan, A. M., & Ragins, B. R. (2017). The evolution and future of diversity at work. *Journal of Applied Psychology, 102*(3), 483–499. https://doi.org/10.1037/apl0000161

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal*, *38*(2), 555–572. https://doi.org/10.5465/256693

Rosenberg, B. D., & Siegel, J. T. (2018). A 50-year review of psychological reactance theory: Do not read this article. *Motivation Science*, *4*(4), 281–300. https://doi.org/10.1037/mot0000091

Rubin, Z., & Peplau, L. A. (1975). Who believes in a just world? *Journal of Social Issues*, *31*(3), 65–89. https://doi.org/10.1111/j.1540-4560.1975.tb00997.x

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, *74*(3), 629–645. https://doi.org/10.1037/0022-3514.74.3.629

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*(4), 743–762. https://doi.org/10.1111/0022-4537.00239

Rynes, S., & Rosen, B. (1995). A field survey of factors affecting the adoption and perceived success of diversity training. *Personnel Psychology*, *48*(2), 247–270. https://doi.org/10.1111/j.1744-6570.1995.tb01756.x

Sanchez, J. I., & Medkik, N. (2004). The effects of diversity awareness training on differential treatment. *Group & Organization Management*, *29*(4), 517–536. https://doi.org/10.1177/1059601103257426

Sidanius, J. (1993). 7. The Psychology of Group Conflict and the Dynamics of Oppression: A Social Dominance Perspective. In S. Iyengar & W. McGuire (Ed.), *Explorations in Political Psychology* (pp. 183–220). Duke University Press. https://doi.org/10.1515/9780822396697-009

Sidanius, J., Levin, S., Federico, C. M., & Pratto, F. (2001). Legitimizing ideologies: The social dominance approach. In J. T. Jost & B. Major (Eds.), *The psychology of legitimacy: Emerging perspectives on ideology, justice, and intergroup relations* (pp. 307–331). Cambridge University Press.

Sidanius, J., & Pratto, F. (1993). Racism and support of free-market capitalism: A cross-cultural analysis. *Political Psychology*, *14*(3), 381–401. https://doi.org/10.2307/3791704

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Sidanius, J., & Pratto, F. (2004). Social dominance theory: A new synthesis. In J. Jost & J. Sidanius (Eds.), *Political Psychology* (pp. 315–332). Psychology Press.

Sidanius, J., Pratto, F., & Bobo, L. (1996). Racism, conservatism, Affirmative Action, and intellectual sophistication: A matter of principled conservatism or group dominance? *Journal of Personality and Social Psychology, 70*(3), 476–490. https://doi.org/10.1037/0022-3514.70.3.476

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93*(2), 280–295. https://doi.org/10.1037/0021-9010.93.2.280

Smith, J. L., McPartlan, P., Poe, J., & Thoman, D. B. (2021). Diversity fatigue: A survey for measuring attitudes towards diversity enhancing efforts in academia. *Cultural Diversity and Ethnic Minority Psychology*, *27*(4), 659–674. https://doi.org/10.1037/cdp0000406

Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal?. *Journal of Vocational Behavior*, *68*(3), 446–460. https://doi.org/10.1016/j.jvb.2005.10.005

Stanhope, D. S., Pond, S. B. III, & Surface, E. A. (2013). Core self-evaluations and training effectiveness: Prediction through motivational intervening mechanisms. *Journal of Applied Psychology*, *98*(5), 820–831. https://doi.org/10.1037/a0032599

Staub, E. (1996). Cultural-societal roots of violence: The examples of genocidal violence and of contemporary youth violence in the United States. *American Psychologist, 51*(2), 117–132. https://doi.org/10.1037/0003-066X.51.2.117

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379–440). Academic Press. https://doi.org/10.1016/S0065-2601(02)80009-0

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance: New developments and findings. *Zeitschrift für Psychologie*, *223*(4), 205–214. https://doi.org/10.1027/2151-2604/a000222

Stewart, R., Volpone, S. D., Avery, D. R., & McKay, P. (2011). You support diversity, but are you ethical? Examining the interactive effects of diversity and ethical climate perceptions on turnover intentions. *Journal of Business Ethics*, *100*(4), 581–593.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Nelson Hall.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology*, *76*(6), 759–769. https://doi.org/10.1037/0021-9010.76.6.759

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*(5), 853–870. https://doi.org/10.1037/0022-3514.78.5.853

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Erlbaum.

Thomas, K. M., & Plaut, V. C. (2008). The many faces of diversity in the workplace. In K. M. Thomas (Ed.), *Diversity resistance in organizations: Manifestations and solutions* (pp. 1–22). Lawrence Erlbaum.

Tran, K. (2021, March 23). The diversity and inclusion industry has lost its way. *Bazaar*. https://www.harpersbazaar.com/culture/features/a35915670/the-diversity-and-inclusion-industry-has-lost-its-way/

Towler, A. J. (2003). Effects of charismatic influence training on attitudes, behavior, and performance. *Personnel Psychology*, *56*(2), 363–381. https://doi.org/10.1111/j.1744-6570.2003.tb00154.x

Voci, A., & Hewstone, M. (2003). Intergroup contact and prejudice toward immigrants in Italy: The mediational role of anxiety and the moderational role of group salience. *Group Processes & Intergroup Relations*, *6*(1), 37–54. https://doi.org/10.1177/1368430203006001011

Wagstaff, G. F. (1983). Correlates of the just world in Britain. *Journal of Social Psychology*, *121*(1), 145–146. https://doi.org/10.1080/00224545.1983.9924476

Ward, A.-K., Beal, D. J., Zyphur, M. J., Zhang, H., & Bobko, P. (2022). Diversity climate, trust, and turnover intentions: A multilevel dynamic system. *Journal of Applied Psychology*, *107*(4), 628–649. https://doi.org/10.1037/apl0000923

Wegner, D. M. (1989). *White bears and other unwanted thoughts: Suppression, obsession, and the psychology of mental control.* Penguin Press.

Wegner, D. M. (1992). You can't always think what you want: Problems in the suppression of unwanted thoughts. In *Advances in experimental social psychology* (pp. 193–225). Academic Press.

Wentling, R. M., & Palma-Rivas, N. (2000). Current status of diversity initiatives in selected multinational corporations. *Human Resource Development Quarterly*, *11*(1), 35–60. https://doi.org/10.1002/1532-1096(200021)11:1<35::AID-HRDQ4>3.0.CO;2-%23

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management*, *17*(3), 601–617. https://doi.org/10.1177/014920639101700305

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: a review of possible mechanisms. *Psychological Bulletin*, *127*(6), 797–826. https://doi.org/10.1037/0033-2909.127.6.797

Yam, K. C., Klotz, A. C., He, W., & Reynolds, S. J. (2017). From good soldiers to psychologically entitled: Examining when and why citizenship behavior leads to deviance. *Academy of Management Journal*, *60*(1), 373–396. https://doi.org/10.5465/amj.2014.0234

Yang, J., & Diefendorff, J. M. (2009). The relations of daily counterproductive workplace behavior with emotions, situational antecedents, and personality moderators: A diary study in Hong Kong. *Personnel Psychology*, *62*(2), 259–295. https://doi.org/10.1111/j.1744-6570.2009.01138.x

Zhang, Y., Fishbach, A., & Dhar, R. (2007). When thinking beats doing: The role of optimistic expectations in goal-based choice. *Journal of Consumer Research*, *34*(4), 567–578. https://doi.org/10.1086/520071

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*(5792), 1451-1452. DOI: 10.1126/science.1130726

Zhong, C.-B., Liljenquist, K., & Cain, D. M. (2009). Moral self-regulation: Licensing and compensation. In D. De Cremer (Ed.), *Psychological perspectives on ethical behavior and decision making* (pp. 75–89). Information Age Publishing.

APPENDIX A TABLE FOR THE CURRENT DEFINITIONS OF DT BACKLASH

Table 1. Current Definitions of DT Backlash

Source	Definition/Description of a DT backlash	Type of responses		
		Negative Emotions	Attitudinal responses	Behavioral responses
Bergman & Salter (2013)	Backlash occurs when people react negatively to policies or decisions that they believe caused others to receive undeserved outcomes		X	
Brannon et al. (2018)	Reactance to policies and practices that are perceived to threaten key goals and motivations		X	
Burke & Black (1997)	Any form of resistance men exhibit toward policies, programs and initiatives undertaken by organizations to promote the hiring and advancement of marginalized employees	X	X	X
Holladay & Quinones (2005, 2008)	*Expected backlash**: Trainees' expectation that DT will increase the prejudice against minority members previous to joining DT		X	
Jones et al. (2013)	A negative response to diversity training that other training programs may not encounter	X	X	
Kaplan (2006)	Employee opposition to diversity, employee's disapproval of other's sexual orientations		X	X
Kidder et al. (2004)	Negative reactions experienced by traditionally higher-status majority group members when they believe that traditionally lower-status minority group members have received preferential treatment; negative reactions to change		X	
Leslie (2019)	A diversity initiative affects intended outcomes or other outcomes in an undesirable directions		X	X

Table 1 (cont'd)

Lindsay (1994)	Reactions of White males who are said to feel over-exposed, targeted and maligned	X		
Mobley & Payne (1992)	Negative reactions to diversity issues	X	X	X
Ragins & Erhardt (2021)	Emotional reactions that can affect training motivation	X		
Sanchez & Medkik (2004)	Negative reactions toward coworkers or supervisors who make the actor participate in DT'	X		

APPENDIX B TABLE FOR THE IDENTIFIED PSYCHOLOGICAL MECHANISMS

Table 2. Psychological Mechanisms of DT Backlash

	Psychological mechanisms	Author Information
Negative emotions	DT may elicit negative emotions (e.g., anger, guilt) for various reasons: 1) when trainees disagree with the contents or DT goals, 2) DT contents and instructions accuse or confront majority group members, or 3) DT facilitators	Kowal (2013), Lindsey et al. (2020)
Psychological reactance	Trainees perceive DT or DT contents as restricting their personal freedom and autonomous motivation, which, in turn, triggers unpleasant motivational arousal	Brannon et al. (2018)
Justice perception	Trainees, especially majority group members (e.g., white males) perceive DT as preferential treatment for the minority group members, therefore DT is an unfair practice	Kidder et al. (2004), Leslie (2019)
Intergroup differences perceptions	The majority group members perceive 1) the contents and messages of DT as a threat to their group's status, 2) some contents of DT (e.g., color-blind approach) excessively highlight the intergroup differences which activate trainees' stereotypes	Pendry et al. (2007), Gundemir et al. (2017)
Moral credentialing	Because participating in DT could be considered moral and ethical, such experiences can morally license trainees, making them perceive it is acceptable to express their prejudice or engage in discriminatory behavior after participating in DT	Leslie (2019)

APPENDIX C TABLES FOR EXPERIMENTAL DESIGNS FOR STUDY 1 AND 2

Table 3. An Experimental Design for Study 1

	Manipulation	Pretraining tests	Training	DT cognitive/ emotional responses	DT attitudinal/behavioral responses
Treatment Group 1	Recall the experiences of prior DT	IAT, justice perceptions	No	Discrete emotions	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Treatment Group 2a/b	Recall the number of prior DT	IAT, justice perceptions	No	Discrete emotions	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Treatment Group 3	Recalling the positiveness of the present company's diversity policies	IAT, justice perceptions	No	Discrete emotions	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Control Group	No manipulation	IAT, justice perceptions	No	Discrete emotions	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW

Table 4. An Experimental Design for Study 2

	Manipulation	Pretraining tests	Training	DT cognitive, emotional responses	DT attitudinal/behavioral responses
Treatment Group 1	Recall the experiences of prior DT	IAT, justice perceptions	Yes	Discrete emotions, Metacognitive activity	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Treatment Group 2a/b	Recall the number of prior DT	IAT, justice perceptions	Yes	Discrete emotions, Metacognitive activity	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Treatment Group 3	Recalling the positiveness of the present company's diversity policiess	IAT, justice perceptions	Yes	Discrete emotions, Metacognitive activity	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW
Control Group 1	No manipulation	IAT, justice perceptions	Yes	Discrete emotions, Metacognitive activity	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW

Table 4. (cont'd)

Control Group 2	No manipulation	IAT, justice perceptions	No	Discrete emotions, Metacognitive activity	IAT, explicit prejudices, willingness to help minority group members, willingness to hire minority members in scenarios, SDO, BJW

APPENDIX D MANIPULATIONS & MATERIALS

Manipulations for Study 1

Manipulation 1

PART 3 will now begin. In this section, you will be asked to recall and write down one of your previous experiences about diversity training.

Then, you will be asked to perform four rounds of implicit association tests.

1. Please think about and write down 2-3 sentences about a particular time when you took diversity training at your company. Did you like it?

2. What did you learn from the diversity training that you just recalled? Please write down 2-3 sentences

3. Did the diversity training that you just recalled influence your work relationship with your colleagues from different backgrounds? Please write down 2-3 sentences.

4. Did the diversity training that you just recalled influence your personal relationship with your colleagues from different backgrounds? Please write down 2-3 sentences.

Manipulation 2

PART 3 will now begin. In this section, you will be asked to recall and write down the number of diversity training you took.

Then, you will be asked to perform four rounds of implicit association tests.

1. Now, please think about the NUMBER of diversity training that you participated in. How many numbers of diversity or DEI-related training did you take when you attended university or college? Please write down the number.

2. How many diversity or DEI-related training did you receive at your current workplace? Please write down the number.

3. How many numbers of diversity or DEI-related training did you receive at your previous workplace? Please write down the number.

4. How many diversity training did you take in total? Please write down the number.

Manipulation 3

PART 3 will now begin. In this section, you will be asked to recall and write down your company's DEI practices or policies.

Then, you will be asked to perform four rounds of implicit association tests.

1. Now, please think about and write down 2-3 sentences about any of your current company’s DEI initiatives or policies

2. Please write down 2-3 sentences about how your company's DEI initiatives or policies helped you to better understand about other minority members or people with different backgrounds.

3. Please write down 2-3 sentences about how your company's DEI initiatives or policies helped people in your workplace to better understand about other minority members or people with different backgrounds

4. Please write down 2-3 sentences about how your company's DEI initiatives or policies helped increased equity and inclusion in the workplace

5. Please write down 2-3 sentences about how your company's DEI initiatives or policies helped people to interact well with each other

Control group

PART 3 will now begin. In this section, you will be asked to perform four rounds of implicit association tests.

Again, it would be greatly appreciated if you could concentrate on the tests.

Manipulations for Study 2

Manipulation 1

PART 3 will now begin. In this section, you will be asked to recall and write down one of your previous experiences about diversity training.

Then, you will watch a short training video on DEI issues.

After watching the video, you will be asked to perform four rounds of implicit association tests.

Please read the following information about diversity training.

A recent study has examined whether people who took diversity training were more moral than people who did not. Regardless of mandatory training or voluntary participation, people who took diversity training showed higher levels of moral reasoning and moral behavior.

1. Please think about and write down 2-3 sentences about a particular time when you took diversity training at your company. Did you like it?

2. Did the diversity training that you just recalled influence your personal relationship with your colleagues from different backgrounds? Please write down 2-3 sentences

Now, you will watch a short video on DEI issues. Please note that you need to answer the questions correctly about the video to receive your incentives.

Manipulation 2

PART 3 will now begin. In this section, you will be asked to recall and write down one of your previous experiences about diversity training.

Then, you will watch a short training video on DEI issues.

After watching the video, you will be asked to perform four rounds of implicit association tests.

Please read the following information about diversity training.

A recent study has examined whether people who took diversity training were more moral than people who did not. Regardless of mandatory training or voluntary participation, people who took diversity training showed higher levels of moral reasoning and moral behavior.

1. How many diversity or DEI-related training did you receive at your current workplace? Please write down the number.

2. How many numbers of diversity or DEI-related training did you receive at your previous workplace? Please write down the number.

3. How many numbers of diversity or DEI-related training did you receive at your previous workplace? Please write down the number.

4. How many diversity training did you take in total? Please write down the number.

Now, you will watch a short training video on DEI issues. Please note that you need to answer the questions correctly about the video to receive your incentives.

Manipulation 3

PART 3 will now begin. In this section, you will be asked to recall and write down your company's DEI policies.

Then, you will watch a short training video on DEI issues.

After watching the video, you will be asked to perform four rounds of implicit association tests.

Please read the following information about diversity policies in a company.

A recent study has examined whether people who endorse diversity policies were more moral than people who do not. Human resources researchers found out that people who support the policies showed higher levels of moral reasoning and moral behavior.

1. Now, please list any of your current company’s DEI initiatives or policies.

2. Please write down 2-3 sentences about how your company’s DEI initiatives or policies helped to maintain equal opportunities at your workplace

3. Please write down 2-3 sentences about how your company’s DEI initiatives or policies helped you to interact well with other people

Now, you will watch a short training video on DEI issues. Please note that you need to answer the questions correctly about the video to receive your incentives.

Control group 1 (who did not recall anything but watched a DT video)

PART 3 will now begin. In this section, you will be asked to watch a short training video on DEI issues.

Then, you will be asked to perform four rounds of implicit association tests.

Now, you will watch a short training video on DEI issues. Please note that you need to answer the questions correctly about the video to receive your incentives.

Control group 2 (who did not recall anything nor watched a DT video)

PART 3 will now begin. In this section, you will be asked to perform four rounds of implicit association tests. Again, it would be greatly appreciated if you could concentrate on the tests.

DT Video Used for Study 2

Royal Bank of Canada. (2020, August 6). When More People Speak Up, More People Listen [Video]. YouTube. https://www.youtube.com/watch?v=b1nJqpqgzR0&t=1s

APPENDIX E MEASURES

Table 5. Measures Used for Study 1 and 2

Psychological State after Manipulations	
Moral credits (5 items) 1. I feel like I have earned recognition for becoming a good person by participating in diversity training 2. My participation in diversity training is evidence that I am a good person 3. Taking diversity training makes me feel like I am a good person 4. Participating in diversity training confirms that I am an ethical person 5. Diversity training offers proof that the participants are ethical	I developed this scale to measure moral credits in the DT context. 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)
Moral credentials (5 items) 1. It would make me feel good to be a person who believes in DEI initiatives. 2. Being someone who supports DEI initiatives is an important part of who I am. 3. I would be ashamed to be a person who does NOT vouch for DEI initiatives. (R) 4. Supporting DEI initiatives is NOT really important to me. (R) 5. I strongly desire to internalize DEI initiatives.	Adapted from Aquino and Reed (2002)'s moral identity internalization 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)
Psychological entitlement (9 items) 1. I honestly feel I'm just more deserving than others. 2. Great things should come to me. 3. If I were on the Titanic, I would deserve to be on the first lifeboat! 4. I demand the best because I'm worth it. 5. I do not necessarily deserve special treatment. 6. I deserve more things in my life. 7. People like me deserve an extra break now and then. 8. Things should go my way. 9. I feel entitled to more of everything.	Campbell et al. (2004) 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)

Table 5 (cont’d)

DT Backlash Manifestations	
DT Backlash Cognitive Manifestations for Study 2	
Metacognitive activity (5 items) 1. I thought about skills I needed to perform what the video was emphasizing 2. I tried to monitor closely the areas where I needed the most improvement 3. I thought about what things I needed to do to learn more 4. I tried to make sure I understood the things from the video 5. I tried to think through the topic and decide what I was supposed to learn from it	Adapted from Schmidt & Ford (2003)’s 15-item scale 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)
DT Backlash Affective Manifestations	
Modern racism scale (MRS) as explicit prejudices (7 items) 1. Discrimination against Blacks is no longer a problem in the United States. 2. Over the past few years, Blacks have gotten more economically than they deserve. 3. It is easy to understand the anger of black people in America. 4. Blacks have more influence upon school desegregation plans than they ought to have. 5. Blacks are getting too demanding in their push for equal rights. 6. Over the past few years the government and news media have shown more respect to blacks than they deserve. 7. Blacks should not push themselves where they're not wanted	McConahay (1986) 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)

Table 5 (cont'd)

Discrete emotions (12 items) 1. Annoyed 2. Irked 3. Stressed 4. Worried 5. Calm 6. Peaceful 7. Grateful 8. Thankful 9. Disappointed 10. Discouraged 11. Confused 12. Puzzled	Chung et al. (2022) 5-point Likert scale (1 = not at all, 5= extremely)
DT Backlash Behavioral Manifestations	
Willingness to help minority members (7 items) 1. I would help minority group member(s) who have been absent. 2. I would minority group member(s) who have heavy work loads. 3. I would assist minority group member(s) with his/her work. 4. I would take time to listen to minority group member's problems and worries. 5. I would go out of my way to help minority group member(s). 6. I would take a personal interest in minority group member(s). 7. I would pass along information to minority group member(s).	Adapted from Williams and Anderson (1991)s' organizational citizenship behavior – Individual to address the willingness to help minority group members. 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)
Hiring decision	Refer to the scenario below this table

Table 5 (cont'd)

Justice Perceptions Moderators	
Distributive justice perceptions (4 items) 1. Promoting DEI initiatives at your organization helps secure outcomes that reflect the effort you have put into your work? 2. Promoting DEI initiatives at your organization is appropriate for your work outcomes? 3. Promoting DEI initiatives at your organization safeguard your contribution to the organization? 4. Promoting DEI initiatives at your organization is fair for your work outcomes given your performance?	Adapted from Colquitt (2001) to reflect DEI values 5-point Likert scale (1= to a small extent, 5= to a large extent)
Procedural justice perception about the assignment of DT (7 items) 1. Have you been able to express your views and feelings when your diversity training was assigned to you? 2. Have you had influence over the diversity training assigned to you at by those procedures? 3. Have those procedures for assigning you to diversity training been applied consistently? 4. Have those procedures for assigning you to diversity training been free of bias? 5. Have those procedures for assigning you to diversity training been based on accurate information? 6. Have you been able to appeal for assigning you to diversity training arrived at by those procedures? 7. Have those procedures upheld ethical and moral standards?	Adapted from Colquitt (2001) to reflect justice perceptions about assignments of DT 5-point Likert scale (1= to a small extent, 5= to a large extent)

Table 5 (cont'd)

Procedural justice perception during DT (7 items) 1. Have you been able to express your views and feelings during diversity training? 2. Have you had influence during the diversity training at by those procedures? 3. Have those procedures during diversity training been applied consistently? 4. Have those procedures during diversity training been free of bias? 5. Have those procedures during diversity training been based on accurate information? 6. Have you been able to appeal the outcomes during diversity training arrived at by those procedures? 7. Have those procedures during diversity training upheld ethical and moral standards?	Adapted from Colquitt (2001) to reflect justice perceptions during DT 5-point Likert scale (1= to a small extent, 5= to a large extent)
Individual Differences Moderators	
Trait psychological reactance (11 items) 1. I become frustrated when I am unable to make free and independent decisions. 2. I become angry when my freedom of choice is restricted. 3. It irritates me when someone points out things which are obvious to me. 4. Regulations trigger a sense of resistance in me. 5. I find contradicting others stimulating. 6. When something is prohibited, I usually think "that's exactly what I am going to do." 7. I resist the attempts of others to influence me. 8. It makes me angry when another person is held up as a model for me to follow. 9. When someone forces me to do something, I feel like doing the opposite. 10. I consider advice from others to be an intrusion. 11. Advice and recommendations induce me to do just the opposite.	Hong & Page (1989) 5-point Likert scale (1 = strongly disagree to 5 = strongly agree)

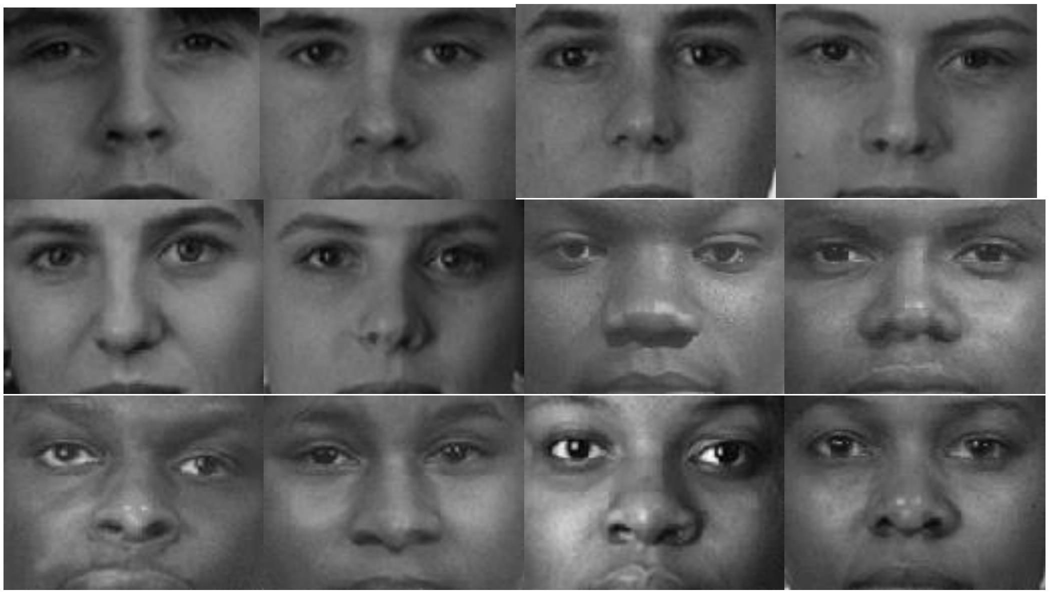
Table 5 (cont'd)

Social dominance orientation (SDO; 8 items) 1. An ideal society requires some groups to be on top and others to be on the bottom. 2. Some groups of people are simply inferior to other groups. 3. No one group should dominate in society. 4. Groups at the bottom are just as deserving as groups at the top. 5. Group equality should not be our primary goal. 6. It is unjust to try to make groups equal. 7. We should do what we can to equalize conditions for different groups. 8. We should work to give all groups an equal chance to succeed.	Ho et al. (2015) 5-point Likert scale (1 = strongly oppose, 5 = strongly favor)
Belief in a just world (BJW; 5 items) 1. Basically, the world is a just place 2. It is a common occurrence for a guilty person to get off free in American courts 3. By and large, people deserve what they get 4. Good deeds often go unnoticed and unrewarded 5. People who meet with misfortune have often brought it on themselves	Original items are from the 20-item scale that Rubin and Peplau (1975) developed. However, the current study followed the shorter version of the original scale following Judge et al. (1998). 5-point Likert scale (1 = strongly disagree, 5 = strongly agree)
Manipulation Checks for Study 1 & 2	
a. What I just asked to recall was…(for manipulation groups 1 & 2) 1. my previous diversity training 2. my previous sexual harassment training b. What I just asked to recall was…(only for manipulation group 3) 1. My company's DEI policies 2. My previous sexual harassment training	Participants who failed to pass a manipulation check were removed from the analyses.

DT Backlash Affective Manifestations – Implicit Association Tests (IAT)

The IAT measures associations between social categories such as gender (e.g., male and female) or race (e.g., European Americans and African Americans) and evaluations (e.g., good, bad). This test requires a computer and a keyboard for participants to complete the sorting tasks. For each round, participants engage in picture-sorting tasks, which come with sorting tasks of positive or negative attributes after seeing the pictures. Every participant goes through some practice rounds. Participants use two buttons ("E" and "I" keys) on a keyboard with both left and right hands to do the sorting tasks. People tend to respond faster when items are more closely related in their minds and are paired with the same button. For example, an implicit preference for European Americans relative to African Americans means that people are faster to sort words when 'European Americans' and 'Good' are paired with the same button relative to when 'African Americans' and 'Good' share a button. Multiple rounds of IAT is administered via Qualtrics survey. The present researcher used IATGEN (http://iatgen.org) to create IAT for Qualtrics and to run the statistical analysis. Below is the pictures and attributes used for IAT.

Figure 1. Pictures Used for IAT in Both Experiments



Positive attributes stimuli: Cheer, enjoy, friend, gentle, happy, heaven, love

Negative attributes stimuli: Damage, evil, gloom, hurt, poison, ugly, vomit

DT Backlash Behavioral Manifestations – Willingness to hire minority group members

Experiment participants read the following scenario about a hiring situation, and then answered on a single-item measure. The scenario and the question are from Monin and Miller (2001).

Imagine that you are the police chief of a small town in a rural area of the U.S. Historically the population of the town has been exclusively White, and attitudes towards other ethnicities tend to be unfavorable. As much as you regret it, you know this is especially the case within your unit. You couldn't help overhearing racist jokes coming from people you otherwise consider excellent officers. In fact, a couple of years aga an African-American patrolman joined your unit, and within a year he quit, complaining about hostile working conditions. You are doing what you can to change attitudes, but your main objective is that the police force should do its job, and so far it has been rather effective so you do not want to provoke any major unrest within the ranks. The time has come to recruit a new officer. As a general rule, officers need to be responsible and trustworthy, show quick intelligence enabling them to make split-second decisions in crisis situations. Recent scandals have also highlighted the need for a high level of integrity, resistance to corruption, mild manners and a calm temper. You have just received applications from the new graduates of the local Police Academy. You wonder whether ethnicity should be a factor in your choice. Do you feel that this specific positions (described above) is better suited for any one ethnicity?

5-point Likert Scale

-2 = Yes, much better for a Black person

0 = No, I do not feel this way at all

+2 = Yes, much better for a White person

APPENDIX F STUDY 1 DESCRIPTIVE STATISTICS & CORRELATION

Table 6. Descriptive Statistics and Zero-order Correlations for Study 1

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Mani. 1	0.25	0.44	(N/A)																	
2. Mani. 2	0.26	0.44	-.34**	(N/A)																
3. Mani. 3	0.24	0.43	-.33*	-.33**	(N/A)															
4. Control	0.25	0.44	-.34**	-.34**	-.33**	(N/A)														
5. Pre-IAT	-1.19	12.41	-.03	-.07	.07	.03	(.87)													
6. Post-IAT	-1.52	13.38	-.01	-.10	.08	.03	.92**	(.84)												
7. Moral credentials	3.38	0.67	-.05	.06	.09	-.10	.08	.09	(.91)											
8. Moral credits	2.84	1.09	-.11*	.01	.10	.01	.05	.04	.46**	(.95)										
9. Psychological entitlement	2.57	0.69	-.02	-.03	.08	-.03	.05	.03	.07	.28**	(.84)									
10. Anger	1.59	0.94	.06	-.03	-.12*	.09	-.05	-.08	-.46**	-.29**	-.03									
11. Explicit prejudice	2.04	0.66	-.01	-.08	-.03	.11*	.03	-.01	-.51**	-.04	.28**	.34**	(.75)							
12. Willingness to help	4.08	0.66	.03	.01	-.05	.01	.04	.07	.52**	.22**	-.11*	-.31**	-.41**	(.90)						
13. Willingness to hire	4.09	1.31	-.06	.01	.11*	-.06	.04	.04	-.07	-.07	.07	-.04	.10*	-.13*	(N/A)					
14. Distributive justice	3.48	1.15	-.14**	-.11*	.03	.00	.04	.04	.66**	.52**	.05	-.35**	-.38**	.39**	-.15**	(.94)				

Table 6 (cont'd)

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
15. Procedural justice – assignment	3.32	0.89	.02	-.02	-.01	.02	.05	.05	.46**	.41**	.02	-.31**	-.22**	.40**	-.14**	.59**	(.84)			
16. Procedural justice – during DT	3.36	0.89	-.01	-.01	-.01	.03	.05	.05	.48**	.42**	.00	-.32**	-.29**	.42**	-.12*	.56**	.89**	(.85)		
17. SDO	2.15	0.70	-.01	-.04	-.02	.06	-.03	-.06	-.55**	-.11*	.22**	.32**	.69**	-.44**	.09	-.40**	-.27**	-.33**	(.94)	
18. BJW	2.59	0.66	-.01	.03	-.01	-.02	.06	.06	-.18**	-.14**	-.04	.08	.09	-.17**	.05	-.13*	-.10	-.11*	.14**	(.58)

Numbers in parentheses represent internal consistency. Because willingness to hire minority group members was a single-item measure, calculating internal consistency was not feasible. Mani. 1 = manipulation group 1, Mani. 2 = manipulation group 2, Mani. 3 = manipulation group 3.

* $p < .05$

** $p < .01$

APPENDIX G STUDY 1 TABLES & FIGURES FOR MAIN EFFECTS OF MANIPULATIONS

Figure 2. ANOVA Results of Manipulations on Anger in Study 1

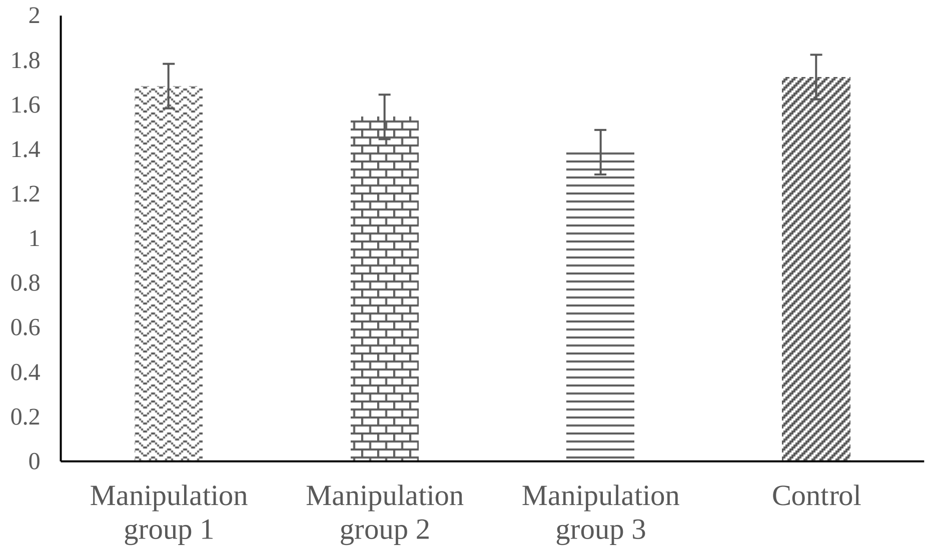


Figure 3. ANOVA Results of Manipulations on Moral Credentials in Study 1

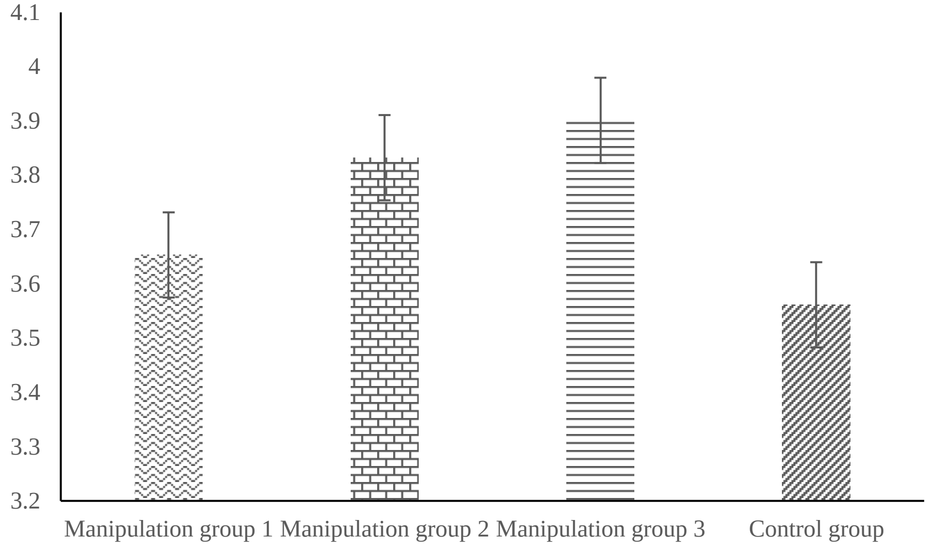
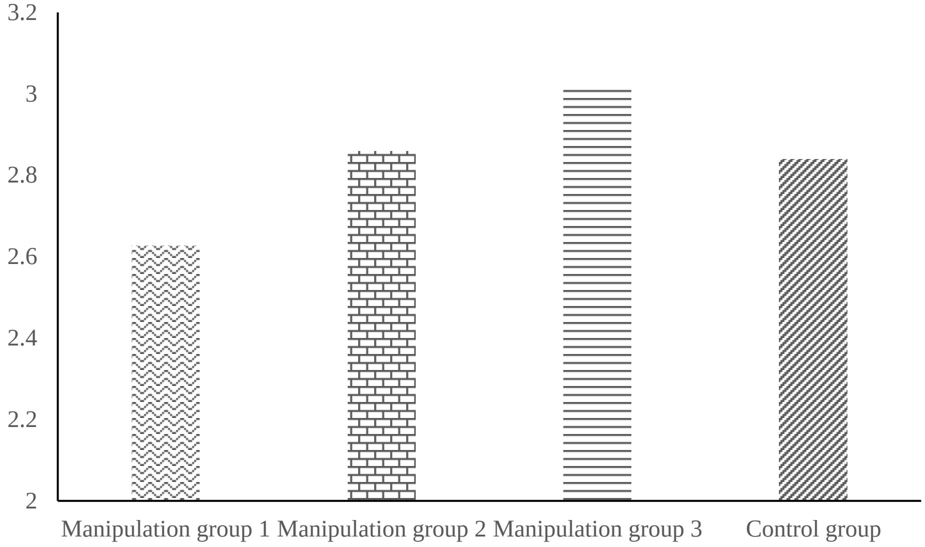


Figure 4. ANOVA Results of Manipulations on Moral Credits in Study 1



APPENDIX H STUDY 1 TABLE FOR MODERATION ANALYSES

Table 7. Results of Direct Effects of Manipulations on DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.05	0.73	0.03	0.15	-0.21	0.15	0.01	0.10	-0.13	0.09	0.03	0.09	-0.02	0.19
Manipulation 2	-1.00	0.73	0.27†	0.15	0.02	0.15	-0.01	0.10	-0.21*	0.09	-0.01	0.09	0.15	0.19
Manipulation 3	0.07	0.74	0.34*	0.15	0.18	0.16	0.13	0.10	-0.16†	0.10	-0.06	0.10	0.37*	0.19
Pre-IAT	1.00**	0.02												
Intercept	-0.12	0.51	3.56**	0.11	2.84**	0.11	2.54**	0.07	2.16**	0.07	4.10**	0.07	3.96**	0.13
R^2	0.85		0.02		0.02		0.01		0.01		0.01		0.01	

[a] *n* = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. Pretest-IAT was entered to test post-test IAT. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members

* $p < .05$

** $p < .01$

† $p < .10$.

Table 8. Results of Moderation Effects of Distributive Justice Perceptions between Manipulations and DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.06	0.74	0.25*	0.11	-0.09	0.13	0.01	0.10	-0.20*	0.09	0.09	0.09	-0.12	0.19
Manipulation 2	-1.12	0.74	0.15	0.11	-0.09	0.13	-0.01	0.10	-0.19*	0.09	-0.05	0.09	0.17	0.18
Manipulation 3	0.07	0.74	0.31**	0.11	0.15	0.13	0.13	0.10	-0.15†	0.09	-0.08	0.09	0.38*	0.19
Pre-IAT	0.99**	0.21												
DJ	0.05	0.47	0.79**	0.07	0.48**	0.09	0.09	0.06	-0.32**	0.06	0.22**	0.06	-0.10	0.12
Manipulation 1 x *DJ*	-0.02	0.62	-0.20*	0.09	-0.02	0.11	-0.13	0.08	0.07	0.07	0.01	0.07	-0.25	0.16
Manipulation 2 x *DJ*	0.49	0.66	-0.23*	0.10	0.05	0.12	-0.06	0.09	0.22**	0.08	0.04	0.08	0.03	0.17
Manipulation 3 x *DJ*	-0.01	0.70	-0.31**	0.11	0.01	0.13	-0.03	0.10	0.14†	0.08	0.01	0.08	-0.06	0.18
Intercept	-0.12	0.52	3.56**	0.08	2.84**	0.09	2.54**	0.07	2.15**	0.06	4.10**	0.06	3.96**	0.13
R^2	0.86		0.46		0.28		0.02		0.17		0.16		0.05	

[a] *n* = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. The moderator was mean-centered. Pretest-IAT was entered to test post-test IAT. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, DJ = Distributive justice. Listwise deletion was used for missing data.

* *p* < .05

** *p* < .01

† *p* < .10.

Table 9. Results of Moderation Effects of Procedural Justice Perceptions about the Assignment of DT between Manipulations and DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.05	0.73	0.11	0.13	-0.20	0.14	0.02	0.10	-0.13	0.09	0.04	0.09	-0.02	0.18
Manipulation 2	-0.99	0.73	0.31*	0.13	0.05	0.14	-0.01	0.10	-0.22*	0.09	0.02	0.09	0.15	0.18
Manipulation 3	0.08	0.74	0.36**	0.13	0.20	0.14	0.13	0.10	-0.16†	0.09	-0.05	0.09	0.37*	0.19
Pre-IAT	0.99**	0.21												
PJa	0.03	0.60	0.63**	0.11	0.51**	0.12	0.05	0.09	-0.19*	0.08	0.37	0.07	-0.12	0.15
Manipulation 1 x *PJa*	-0.01	0.79	-0.02	0.14	-0.13	0.15	-0.17	0.11	-0.07	0.10	-0.06	0.09	-0.27	0.20
Manipulation 2 x *PJa*	0.32	0.83	-0.08	0.15	0.06	0.16	0.01	0.11	0.09	0.10	-0.12	0.10	0.10	0.21
Manipulation 3 x *PJa*	0.03	0.93	-0.33†	0.17	0.14	0.18	0.07	0.13	0.11	0.12	-0.16	0.11	-0.15	0.23
Intercept	-0.12	0.52	3.54**	0.09	2.82**	0.11	2.54**	0.07	2.16**	0.07	4.09**	0.06	3.96**	0.13
R^2	0.85		0.24		0.19		0.02		0.08		0.17		0.04	

[a] n = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. The moderator was mean-centered. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, PJa = Procedural justice regarding the assignment of DT.

* $p < .05$
** $p < .01$
† $p < .10$.

Table 10. Results of Moderation Effects of Procedural Justice Perceptions during DT between Manipulations and DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.05	0.73	0.14	0.13	-0.18	0.14	0.01	0.10	-0.15	0.09	0.05	0.09	-0.03	0.19
Manipulation 2	-1.00	0.73	0.31*	0.13	0.02	0.14	-0.01	0.10	-0.23*	0.09	0.02	0.09	0.15	0.18
Manipulation 3	0.08	0.74	0.38**	0.13	0.22	0.14	0.13	0.10	-0.17†	0.09	-0.04	0.09	0.37*	0.19
Pre-IAT	0.99**	0.21												
PJd	0.01	0.59	0.73**	0.10	0.55**	0.11	0.04	0.08	-0.32**	0.07	0.40**	0.07	-0.03	0.15
Manipulation 1 x *PJd*	0.05	0.78	-0.10	0.14	-0.13	0.15	-0.13	0.11	0.05	0.10	-0.10	0.10	-0.27	0.20
Manipulation 2 x *PJd*	0.43	0.83	-0.17	0.15	0.02	0.16	0.02	0.11	0.20*	0.10	-0.08	0.10	-0.04	0.21
Manipulation 3 x *PJd*	0.07	0.90	-0.46**	0.16	-0.01	0.17	0.01	0.12	0.19*	0.11	-0.23*	0.11	-0.27	0.23
Intercept	-0.12	0.52	3.53**	0.09	2.81**	0.10	2.54**	0.07	2.17**	0.07	4.08**	0.07	3.96**	0.13
R^2	0.85		0.27		0.20		0.01		0.12		0.19		0.04	

^{a} *n* = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. The moderator was mean-centered. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, PJd = Procedural justice during DT.

* $p < .05$

** $p < .01$

† $p < .10$.

Table 11. Results of Moderation Effects of SDO between Manipulations and DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	B	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.05	0.73	0.04	0.13	-0.23	0.16	0.02	0.10	-0.08	0.07	0.01	0.09	0.01	0.19
Manipulation 2	-1.11	0.73	0.18	0.13	0.01	0.15	0.02	0.10	-0.14*	0.07	-0.05	0.09	0.17	0.19
Manipulation 3	0.07	0.74	0.27	0.13	0.16	0.16	0.14	0.10	-0.09	0.07	-0.10	0.10	0.40*	0.19
Pre-IAT	0.99**	0.21												
SDO	-0.02	0.67	-0.72**	0.11	-0.21	0.14	0.07	0.09	0.68**	0.06	-0.38**	0.08	0.24	0.17
Manipulation 1 x *SDO*	-0.08	0.99	-0.24	0.17	-0.01	0.21	0.14	0.13	-0.01	0.09	-0.13	0.12	-0.21	0.25
Manipulation 2 x *SDO*	-2.05*	1.02	-0.12	0.18	0.20	0.22	0.28*	0.14	-0.09	0.10	-0.04	0.12	-0.18	0.26
Manipulation 3 x *SDO*	-0.08	1.06	-0.02	0.18	0.03	0.23	0.26†	0.14	0.01	0.10	0.02	0.12	0.17	0.27
Intercept	-0.12	0.51	3.61**	0.09	2.85**	0.11	2.54**	0.07	2.11**	0.05	4.13**	0.06	3.94**	0.13
R^2	0.86		0.32		0.03		0.07		0.49		0.20		0.03	

[a] *n* = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. The moderator was mean-centered. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, PJd = Procedural justice during DT.

* $p < .05$

** $p < .01$

† $p < .10$.

Table 12. Results of Moderation Effects of BJW between Manipulations and DT Backlash Manifestations for Study 1

	DT backlash manifestations													
DVs	Post-IAT		MCD		MC		Entitle		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.05	0.73	0.10	0.15	-0.21	0.15	0.01	0.10	-0.13	0.09	0.03	0.09	-0.02	0.19
Manipulation 2	-1.02	0.73	0.29^{*}	0.15	0.03	0.15	-0.01	0.10	-0.22^{*}	0.09	0.01	0.09	0.15	0.19
Manipulation 3	0.07	0.74	0.35^{*}	0.15	0.18	0.16	0.13	0.10	$-0.16^{\dagger}$	0.10	-0.06	0.09	0.38^{*}	0.19
Pre-IAT	0.99^{**}	0.21												
BJW	0.06	0.88	-0.56^{*}	0.18	-0.12	0.19	0.15	0.12	0.33^{**}	0.11	-0.16	0.12	0.01	0.22
Manipulation 1 x *BJW*	-0.20	1.15	0.14	0.23	0.01	0.24	-0.14	0.16	$-0.27^{\dagger}$	0.15	-0.12	0.15	0.11	0.29
Manipulation 2 x *BJW*	0.49	1.15	0.51^{*}	0.23	-0.15	0.24	-0.30	0.16	-0.30^{*}	0.15	0.05	0.15	0.08	0.29
Manipulation 3 x *BJW*	-0.08	1.22	0.33	0.24	-0.29	0.26	$-0.25^{\dagger}$	0.16	-0.33^{*}	0.16	0.05	0.16	0.17	0.31
Intercept	-0.12	0.52	3.55^{**}	0.10	2.84^{**}	0.11	2.55^{**}	0.07	2.16^{**}	0.07	4.10^{**}	0.07	3.96^{**}	0.13
R^2	0.85		0.07		0.04		0.02		0.04		0.04		0.02	

[a] n = 387. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was the control group. The moderator was mean-centered. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, PJd = Procedural justice during DT.

* $p < .05$

** $p < .01$

† $p < .10$.

APPENDIX I STUDY 1 FIGURES FOR MODERATION ANALYSES

Figure 5. Interaction between Manipulation 1 and Distributive Justice on Moral Credentials (Study 1)

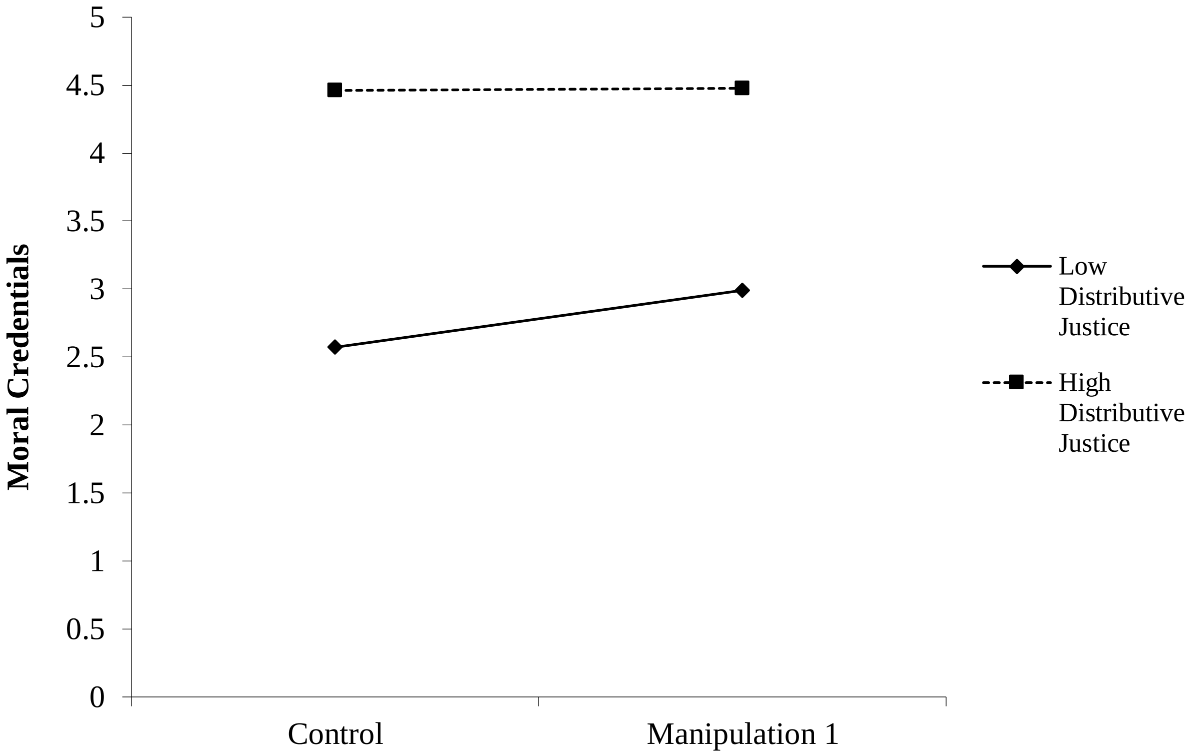


Figure 6. Interaction between Manipulation 2 and Distributive Justice on Moral Credentials (Study 1)

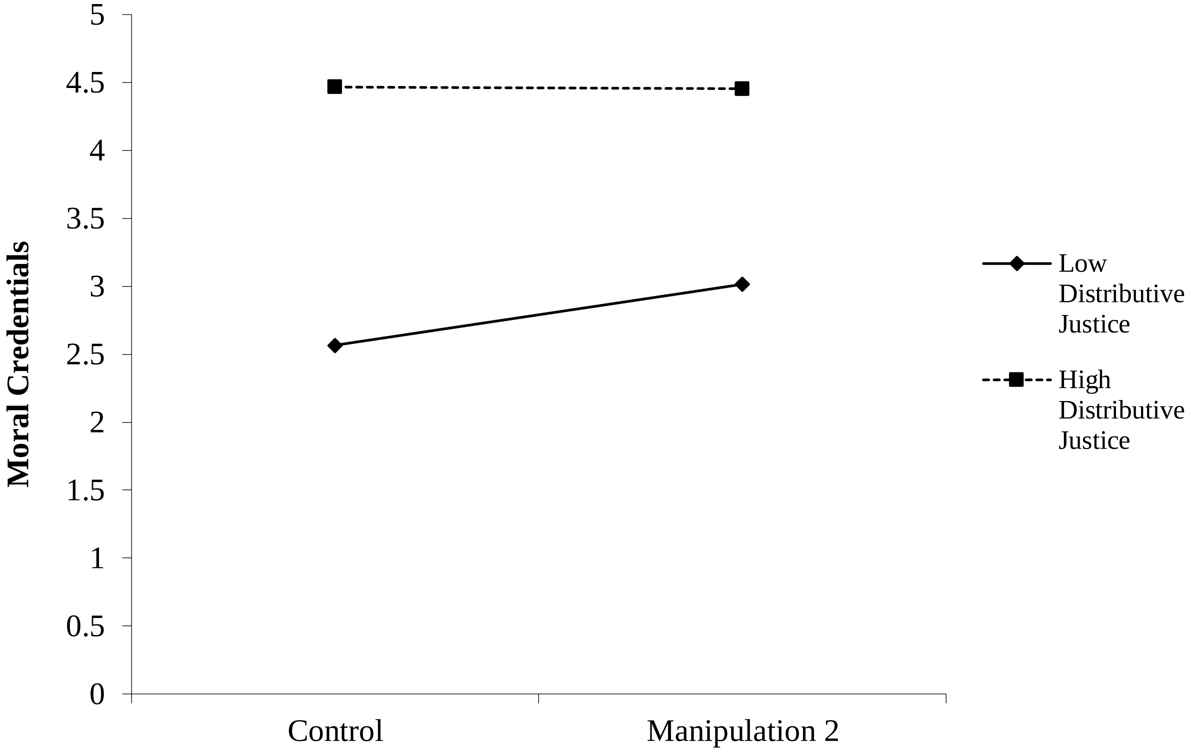


Figure 7. Interaction between Manipulation 3 and Distributive Justice on Moral Credentials (Study 1)

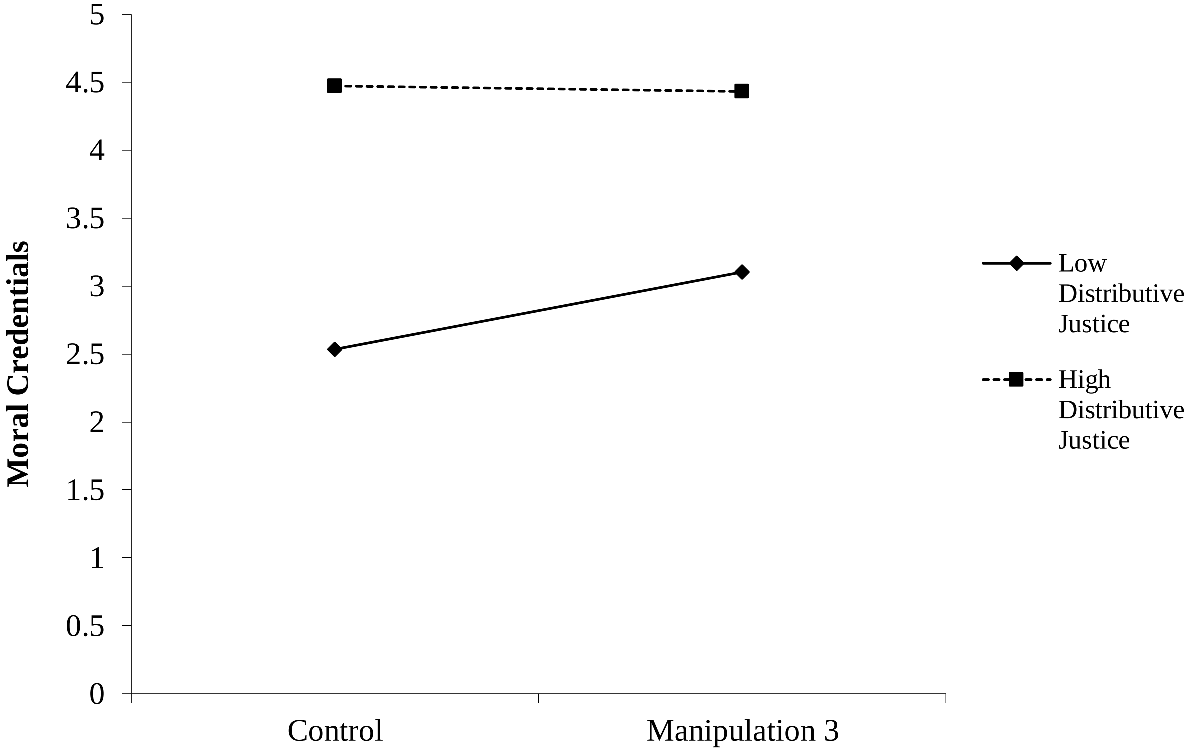


Figure 8. Interaction between Manipulation 1 and Distributive Justice on Anger (Study 1)

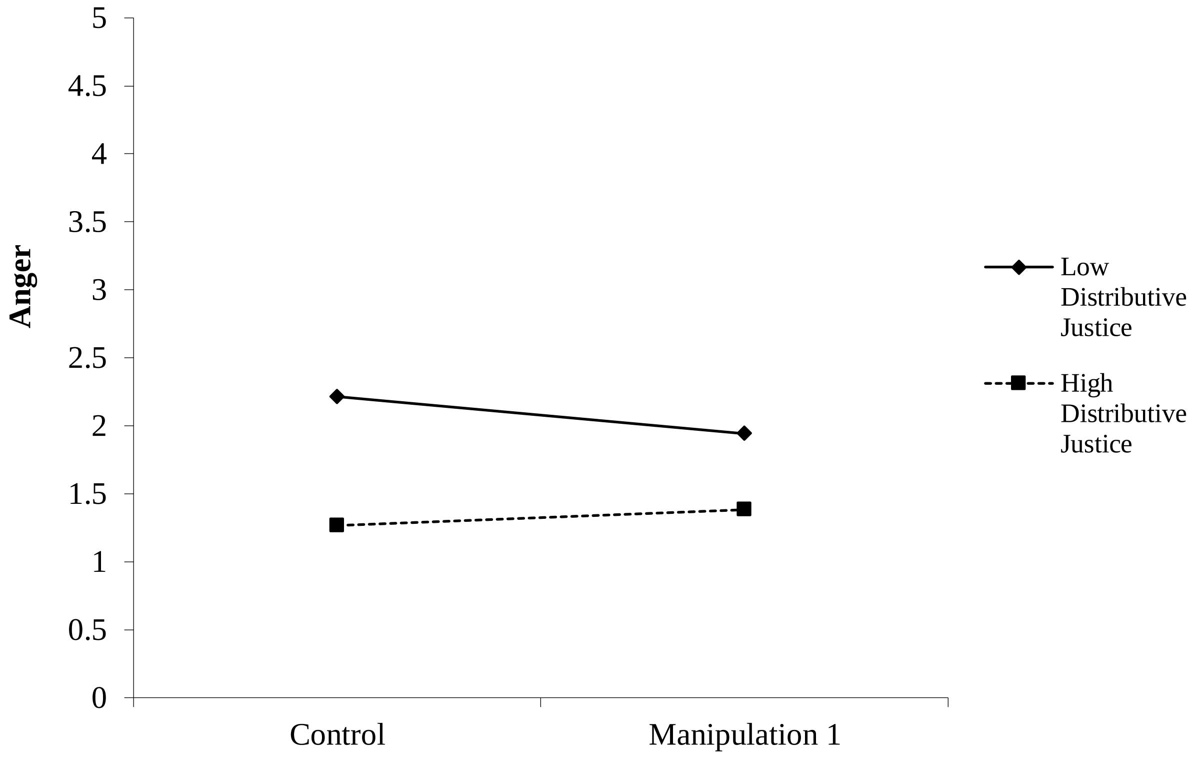


Figure 9. Interaction between Manipulation 2 and Distributive Justice on Explicit Prejudice (Study 1)

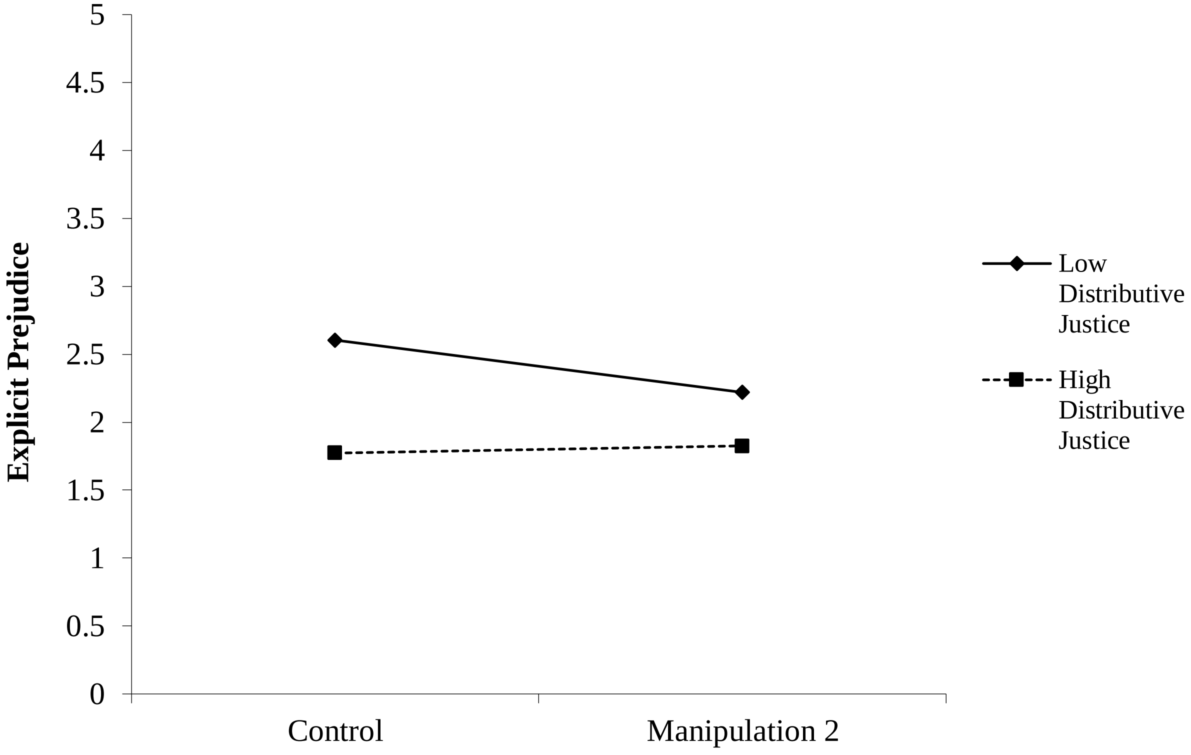


Figure 10. Interaction between Manipulation 3 and Distributive Justice on Explicit Prejudice (Study 1)

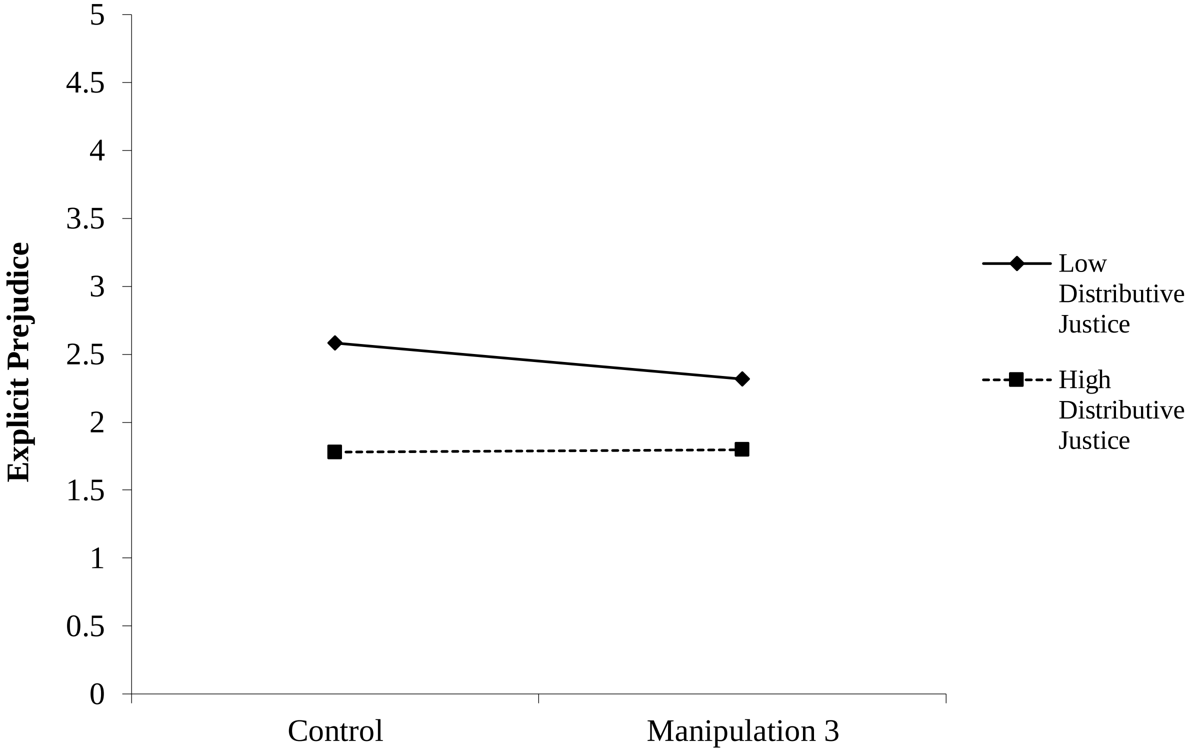


Figure 11. Interaction between Manipulation 3 and Procedural Justice regarding Assignment of DT on Moral Credentials (Study 1)

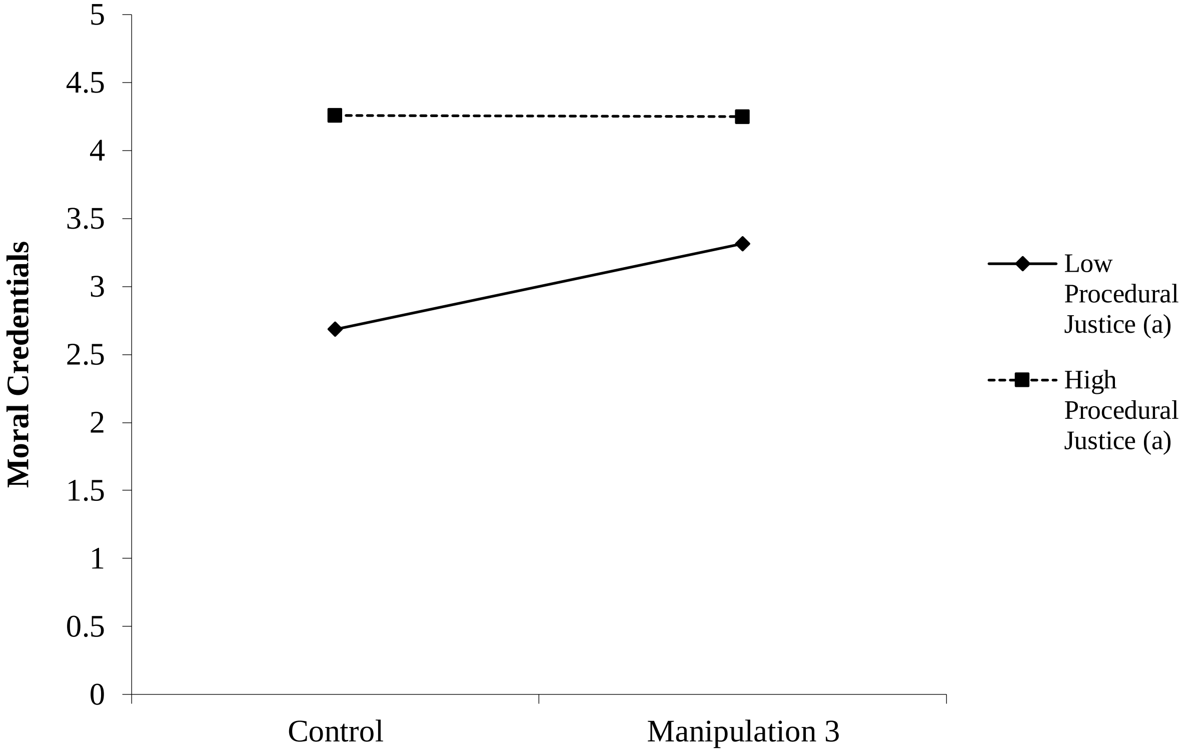


Figure 12. Interaction between Manipulation 3 and Procedural Justice during DT on Moral Credentials (Study 1)

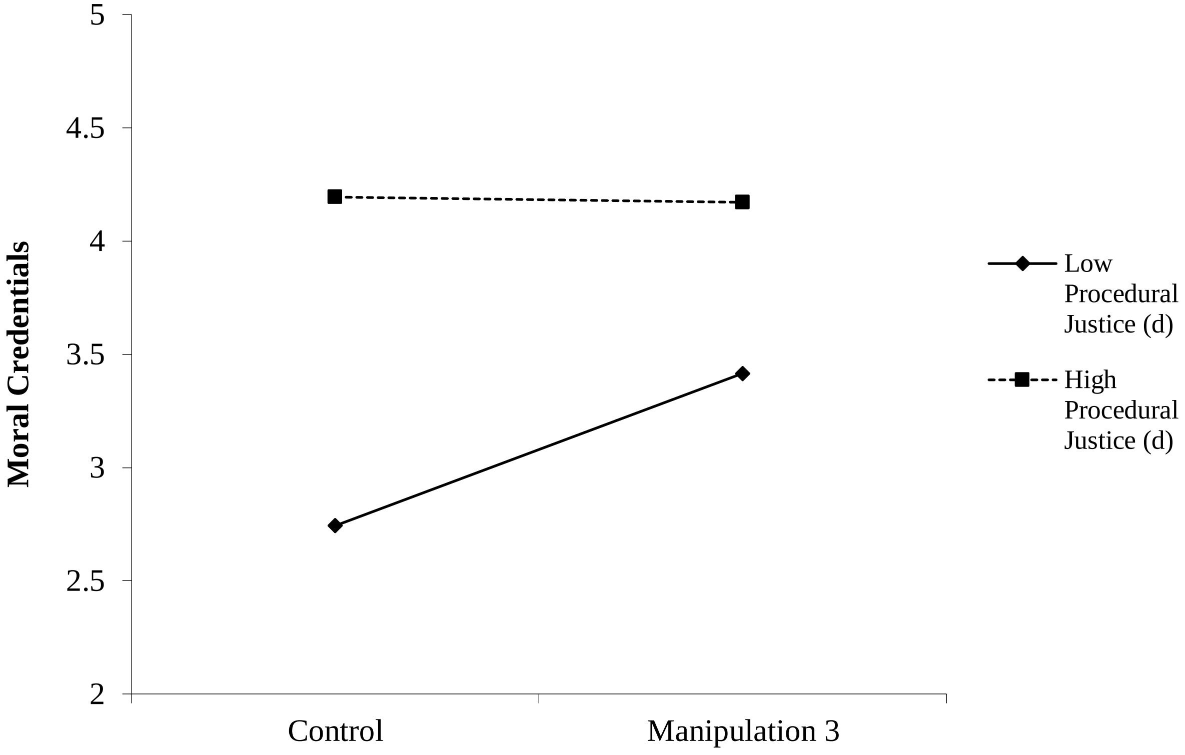


Figure 13. Interaction between Manipulation 2 and Procedural Justice during DT on Explicit Prejudice (Study 1)

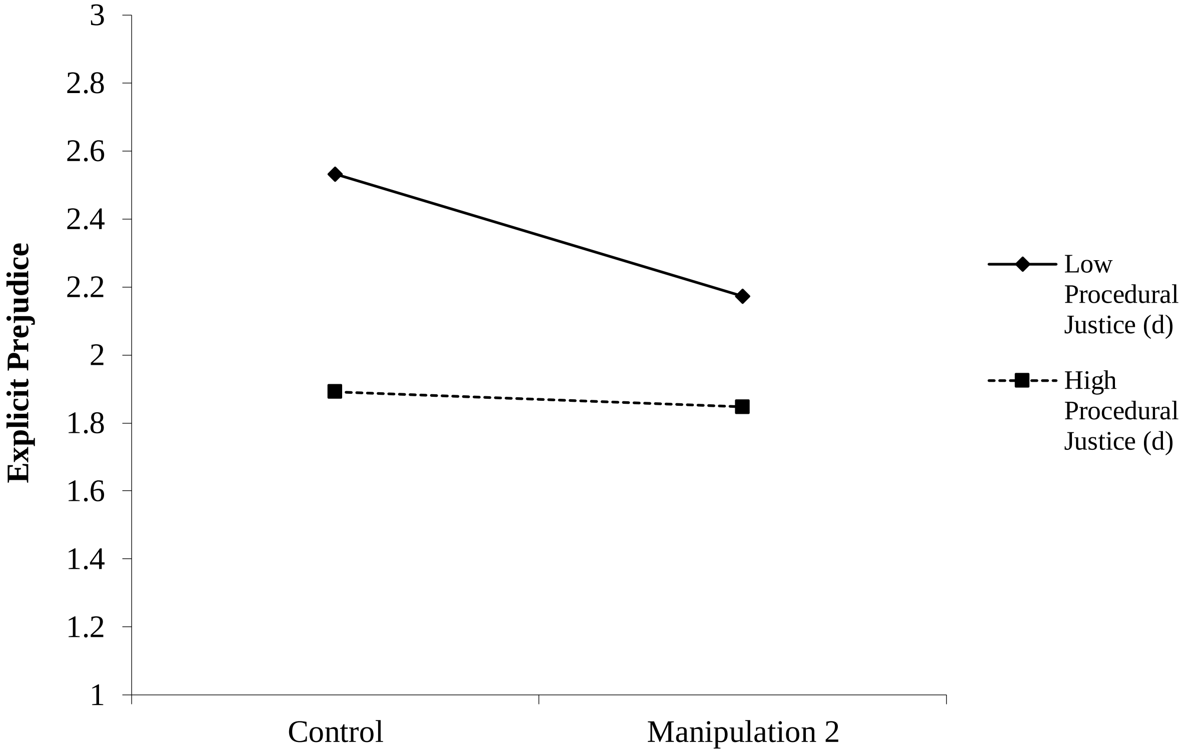


Figure 14. Interaction between Manipulation 3 and Procedural Justice during DT on Explicit Prejudice (Study 1)

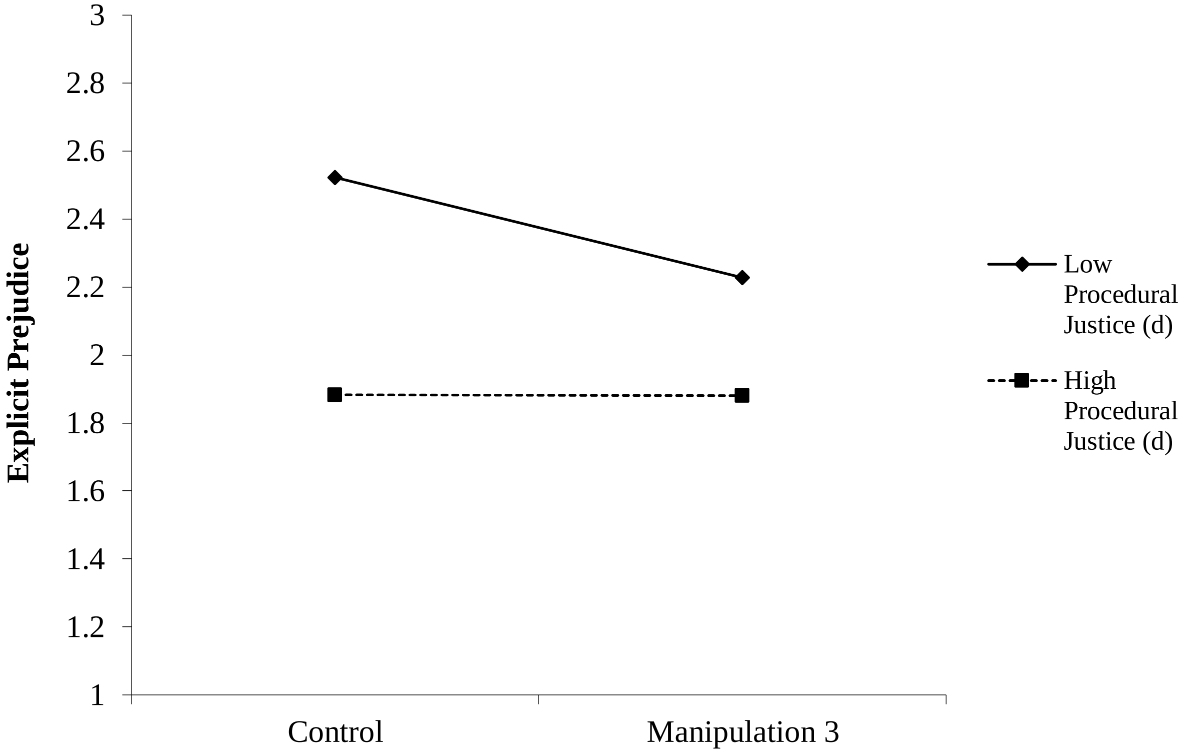


Figure 15. Interaction between Manipulation 3 and Procedural Justice during DT on OCB (Study 1)

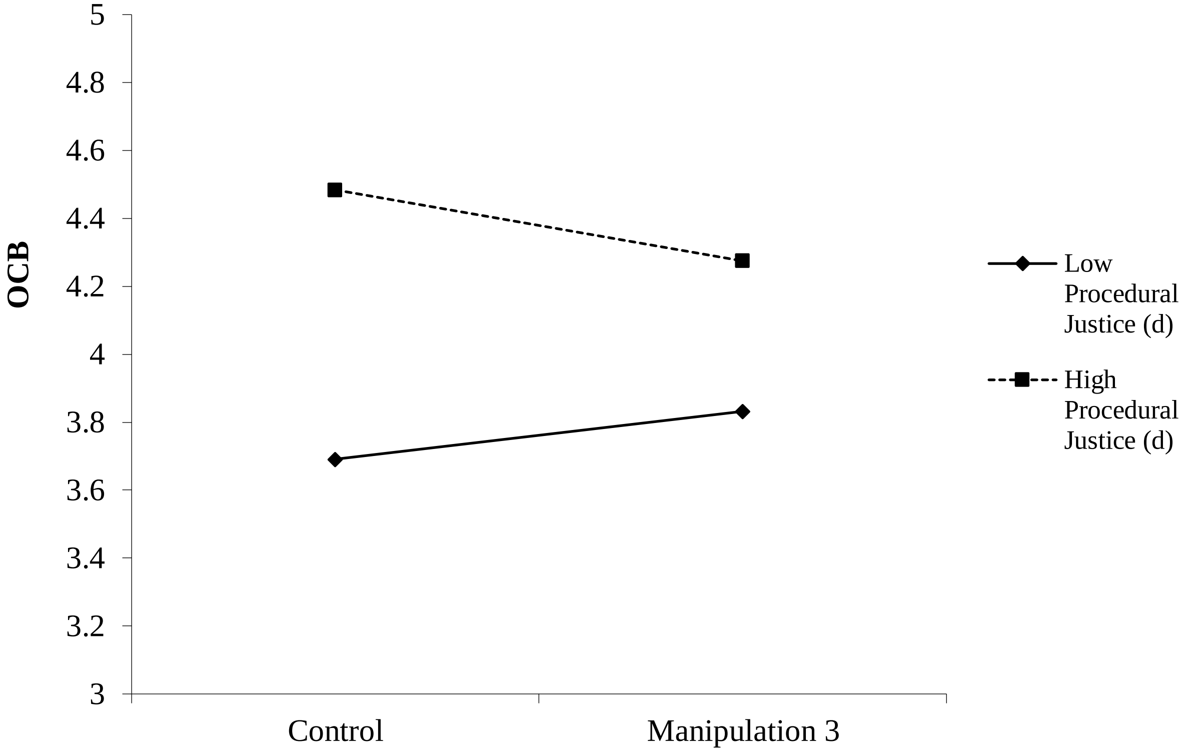


Figure 16. Interaction between Manipulation 2 and SDO on Implicit Prejudice (IAT score; Study 1)

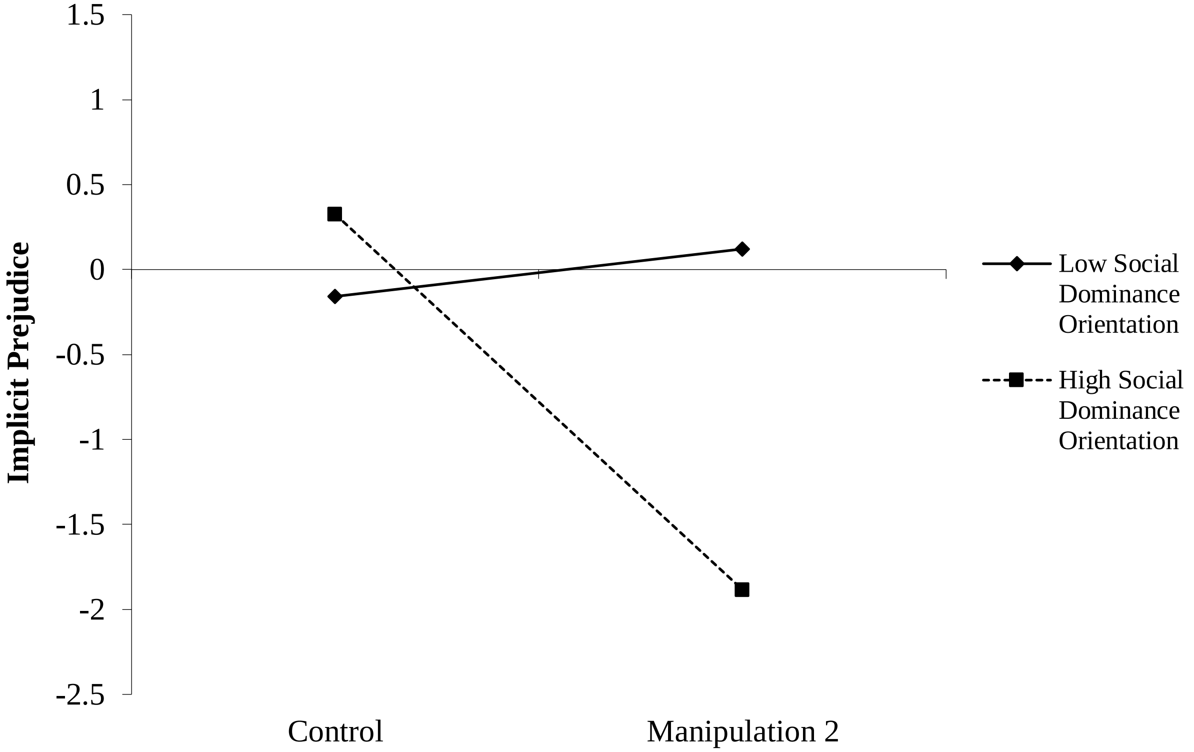


Figure 17. Interaction between Manipulation 2 and SDO on Psychological Entitlement (Study 1)

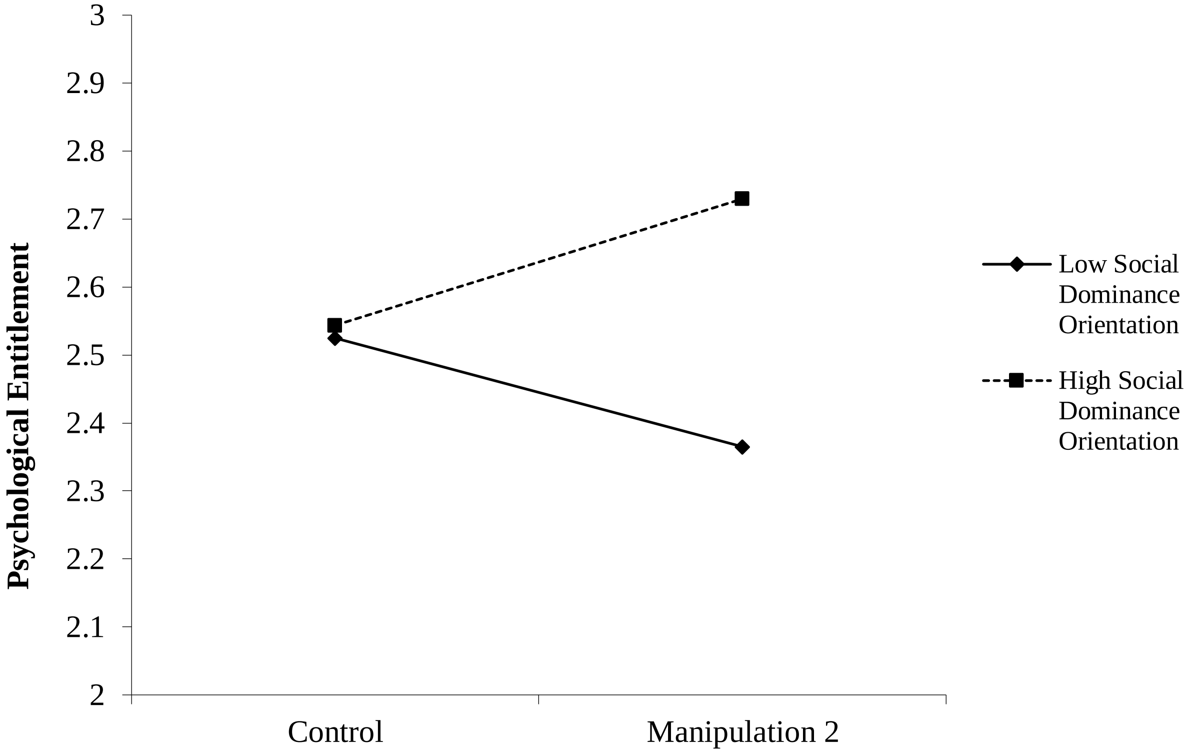


Figure 18. Interaction between Manipulation 3 and SDO on Psychological Entitlement (Study 1)

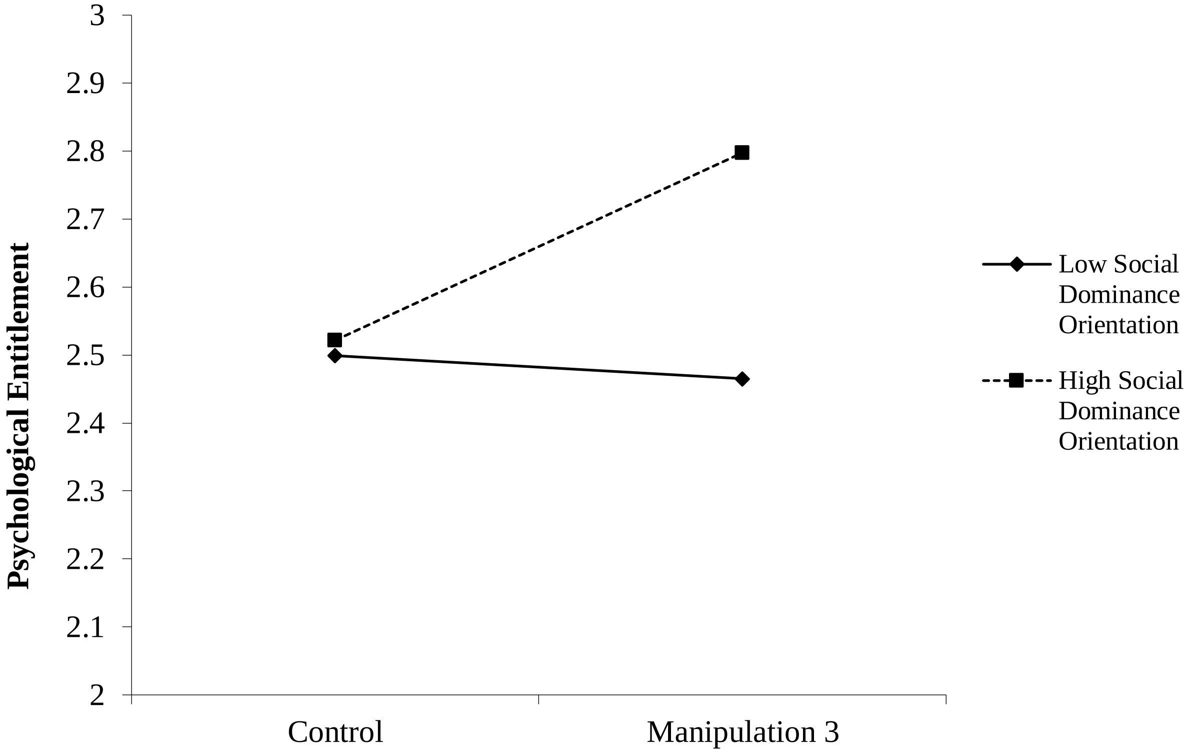


Figure 19. Interaction between Manipulation 2 and BJW on Moral Credentials (Study 1)

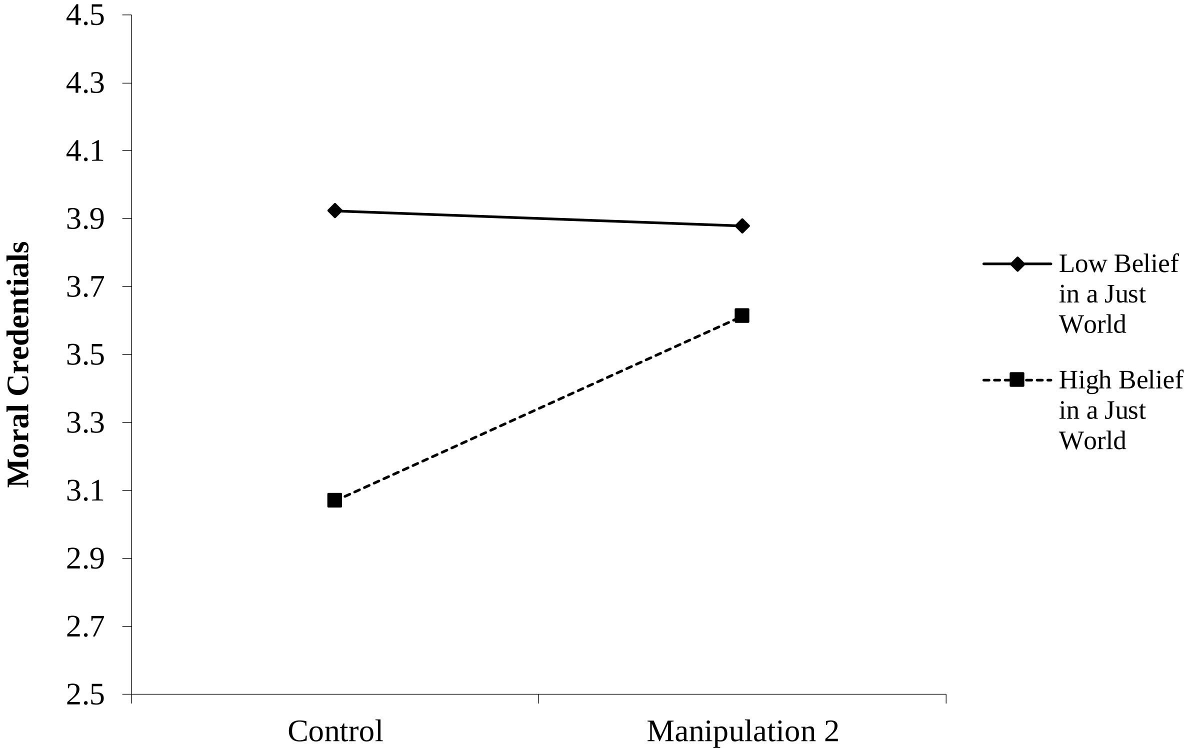


Figure 20. Interaction between Manipulation 2 and BJW on Psychological Entitlement (Study 1)

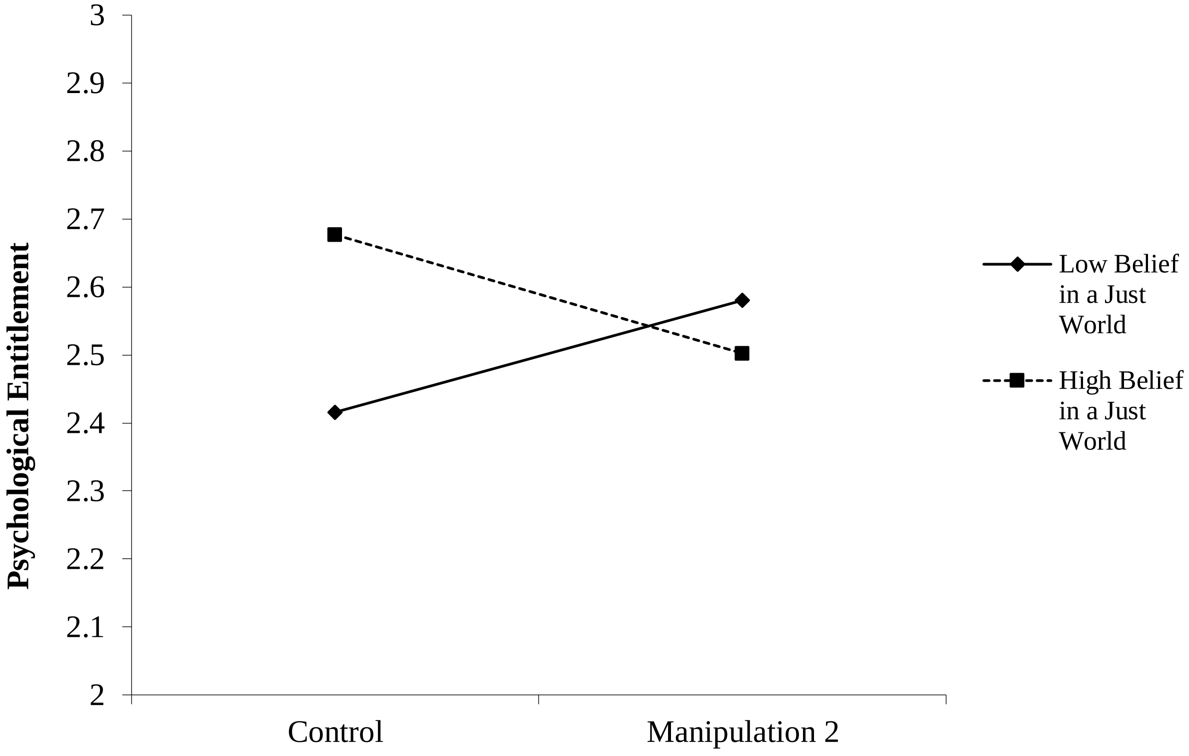


Figure 21. Interaction between Manipulation 1 and BJW on Explicit Prejudice (Study 1)

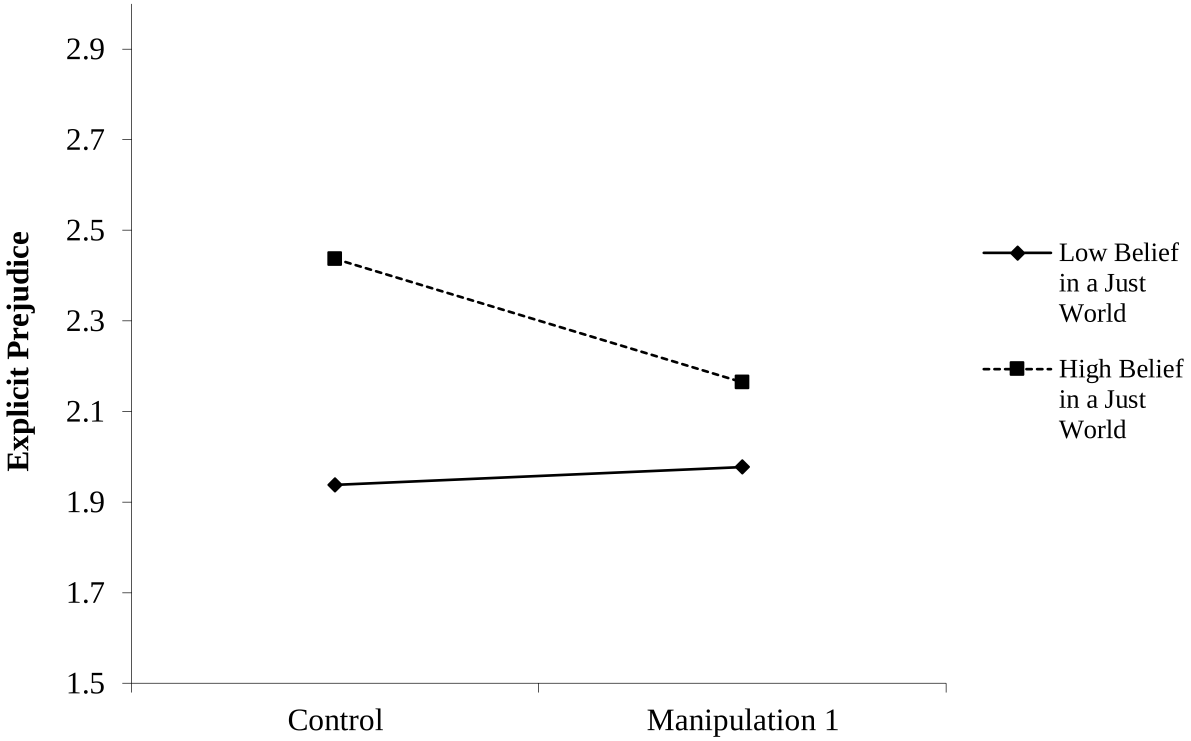


Figure 22. Interaction between Manipulation 2 and BJW on Explicit Prejudice (Study 1)

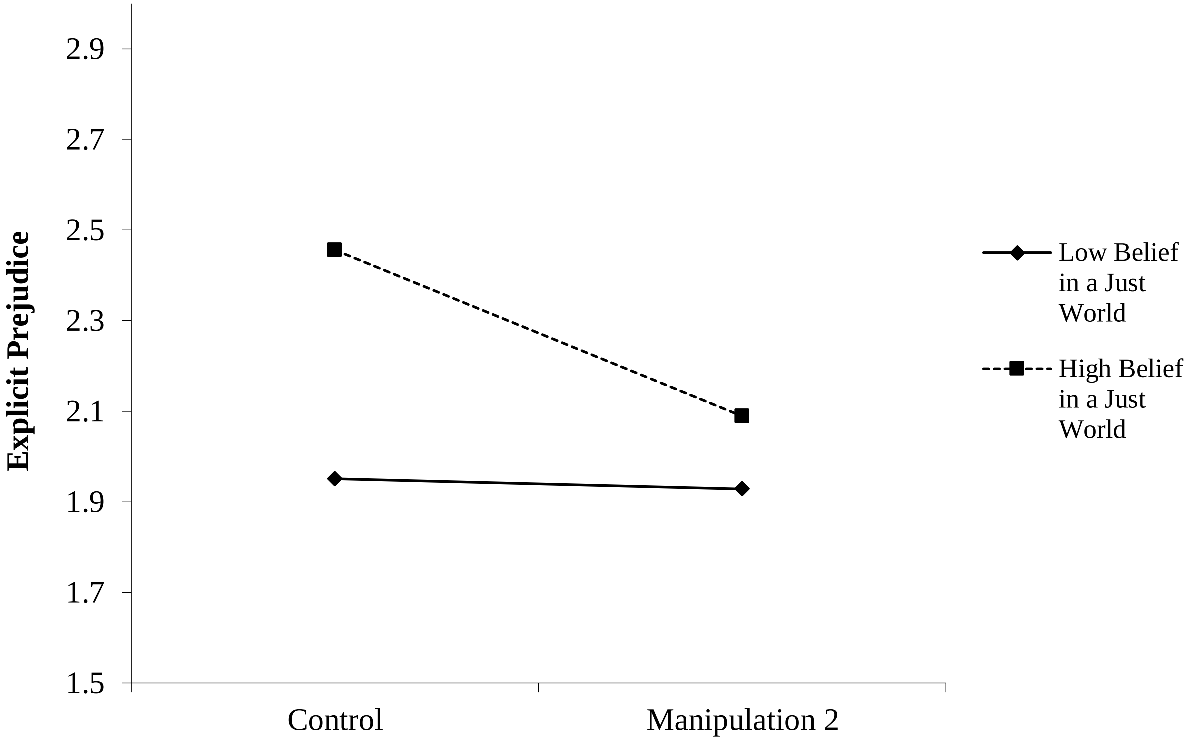
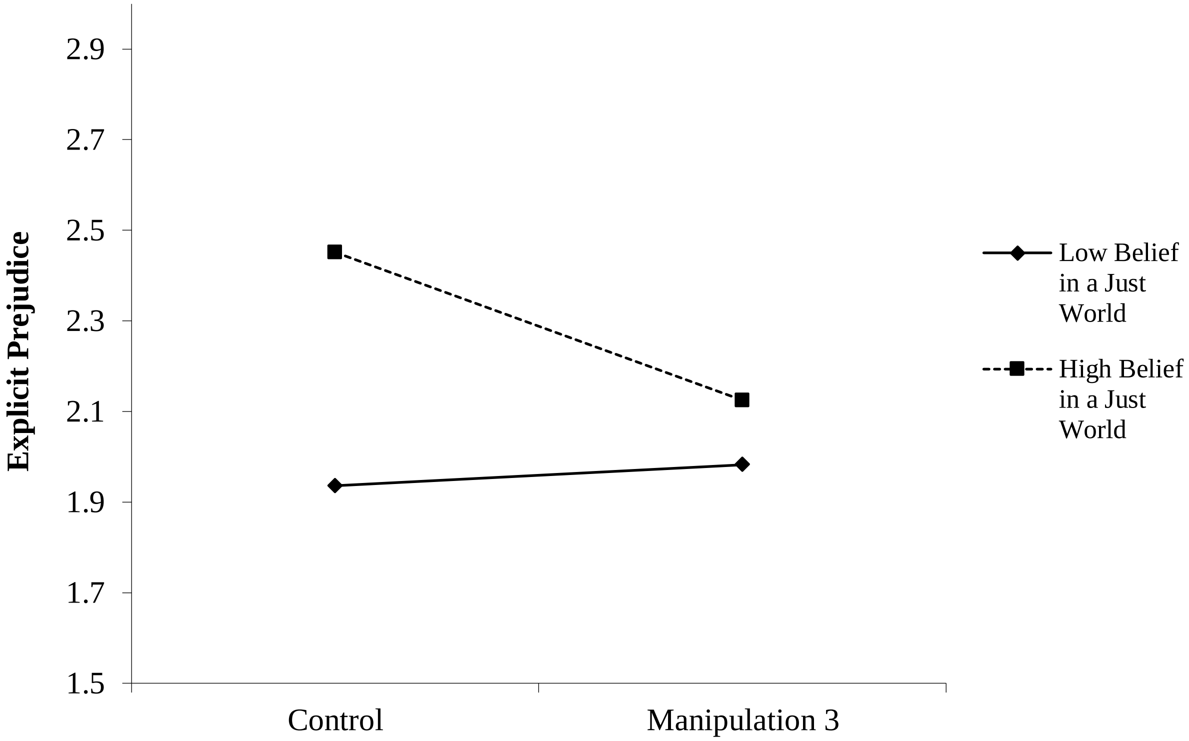


Figure 23. Interaction between Manipulation 3 and BJW on Explicit Prejudice (Study 1)



APPENDIX J STUDY 2 DESCRIPTIVE STATISTICS & CORRELATION

Table 13. Descriptive Statistics and Zero-order Correlations for Study 2

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Mani 1	0.19	0.40	(N/A)																		
2. Mani 2	0.20	0.40	$-.25^{**}$	(N/A)																	
3. Mani 3	0.19	0.39	$-.24^{**}$	$-.24^{**}$	(N/A)																
4. Control 1	0.20	0.40	$-.24^{**}$	$-.25^{**}$	$-.24^{**}$	(N/A)															
5. Control 2	0.22	0.41	$-.26^{**}$	$-.27^{**}$	$-.25^{**}$	$-.26^{**}$	(N/A)														
6. Pre-IAT	-1.33	13.06	.03	-.02	-.02	.07	-.05	(.87)													
7. Post-IAT	-1.81	14.56	.04	-.03	.03	.00	-.03	$.67^{**}$	(.84)												
8. Moral credentials	3.96	0.94	.04	-.11	.11	.05	-.08	.02	.08	(.90)											
9. Moral credits	2.93	1.08	-.03	-.03	.05	-.01	.01	-.09	-.05	$.42^{**}$	(.95)										
10. Psychological entitlement	2.55	0.68	-.05	-.01	-.01	.03	.02	$-.10^{*}$	$-.19^{**}$	-.07	$.18^{**}$	(.83)									
11. Metacognitive activity	3.93	0.78	$.13^{**}$	$-.10^{*}$	.07	.03	$-.12^{*}$	-.01	.02	$.58^{**}$	$.42^{**}$	-.06	(.88)								
12. Anger	1.31	0.77	-.05	.07	-.09	-.05	$.11^{*}$	-.04	$-.13^{**}$	$-.52^{**}$	$-.28^{**}$	$.13^{**}$	$-.39^{**}$	(.84)							
13. Explicit prejudice	2.02	0.68	-.05	.05	-.07	-.03	.09	-.08	$-.21^{**}$	$-.52^{**}$	$-.10^{*}$	$.30^{**}$	$-.28^{**}$	$.37^{**}$	(.78)						
14. Willingness to help	4.11	0.72	.01	-.09	.09	.02	-.03	.05	$.17^{**}$	$.54^{**}$	$.25^{**}$	$-.23^{**}$	$.44^{**}$	$-.34^{**}$	$-.51^{**}$	(.93)					
15. Willingness to hire	4.08	1.19	.10	$-.11^{*}$	-.03	.09	.03	-.01	.01	$-.09^{*}$	.02	.08	-.05	.06	$.13^{**}$	-.06	(N/A)				
16. Distributive justice	3.73	1.14	.00	$-.12^{*}$	.08	-.01	.04	-.08	.01	$.68^{**}$	$.40^{**}$	-.01	$.47^{**}$	$-.39^{**}$	$-.44^{**}$	$.46^{**}$	-.01	(.95)			

Table 13 (cont'd)

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
17. Procedural justice – assignment	3.37	0.93	.04	-.13**	.11**	.01	-.02	-.07	-.02	.50**	.45**	.05	.43**	-.37**	-.23**	.34**	.02	.59**	(.86)		
18. SDO	2.07	0.66	-.08	.08	-.03	-.02	.03	-.08	-.16**	-.58**	-.18**	.22**	-.31**	.39**	.67**	-.51**	.09	-.50**	-.27**	(.77)	
19. BJW	2.65	0.65	-.05	.01	.02	.08	-.05	-.09	-.13**	-.16**	.21**	.18**	.05	.05	.35**	-.17**	.11*	-.13**	.09	.36**	(.59)

Numbers in parentheses represent internal consistency. Because willingness to hire minority group members was a single-item measure, calculating internal consistency was not feasible.

* $p < .05$

** $p < .01$

APPENDIX K STUDY 2 TABLES & FIGURES FOR MAIN EFFECTS OF MANIPULATIONS

Figure 24. ANOVA Results of Manipulations on Anger in Study 2

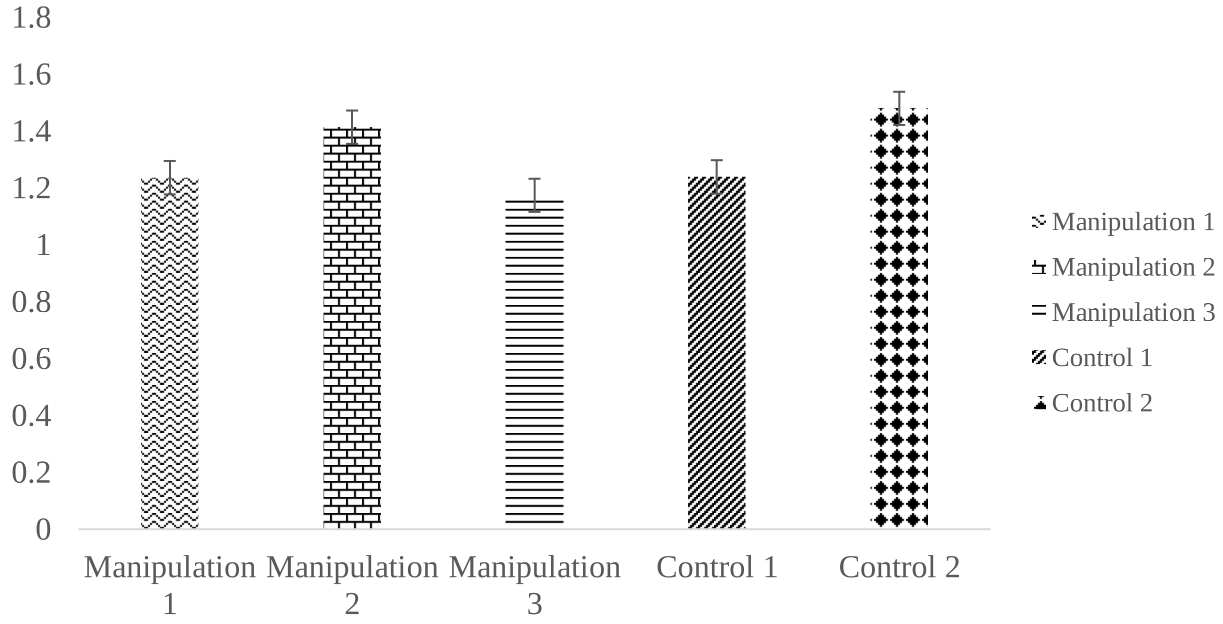


Figure 25. ANOVA Results of Manipulations on Moral Credentials in Study 2

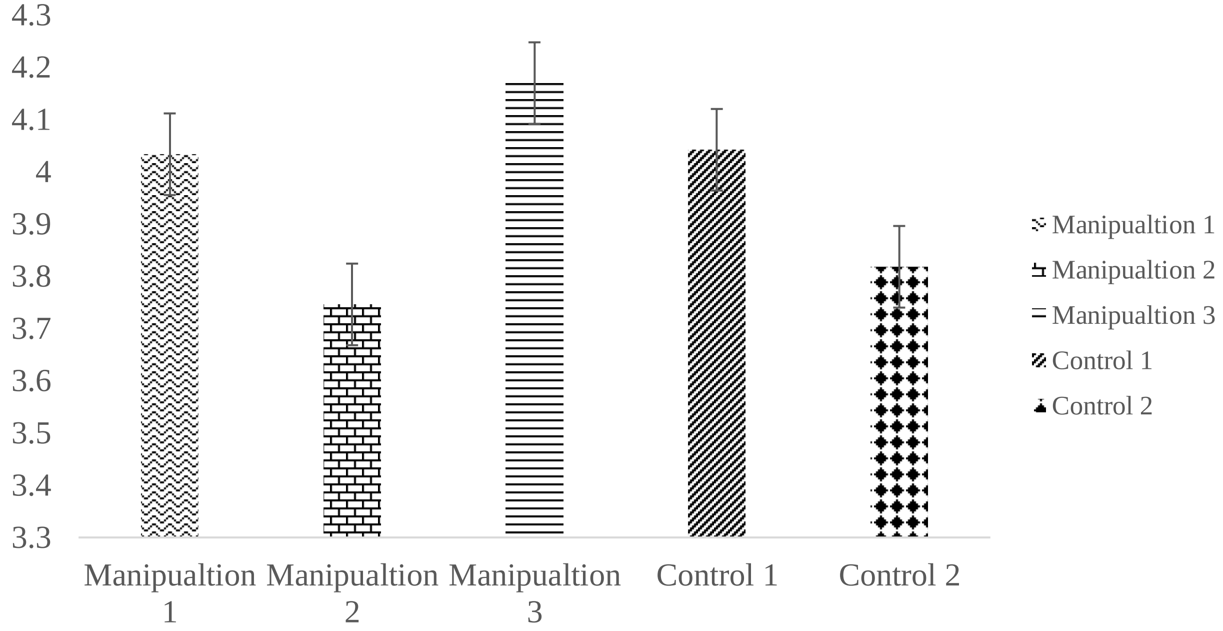


Figure 26. ANOVA Results of Manipulations on Metacognitive Activity in Study 2

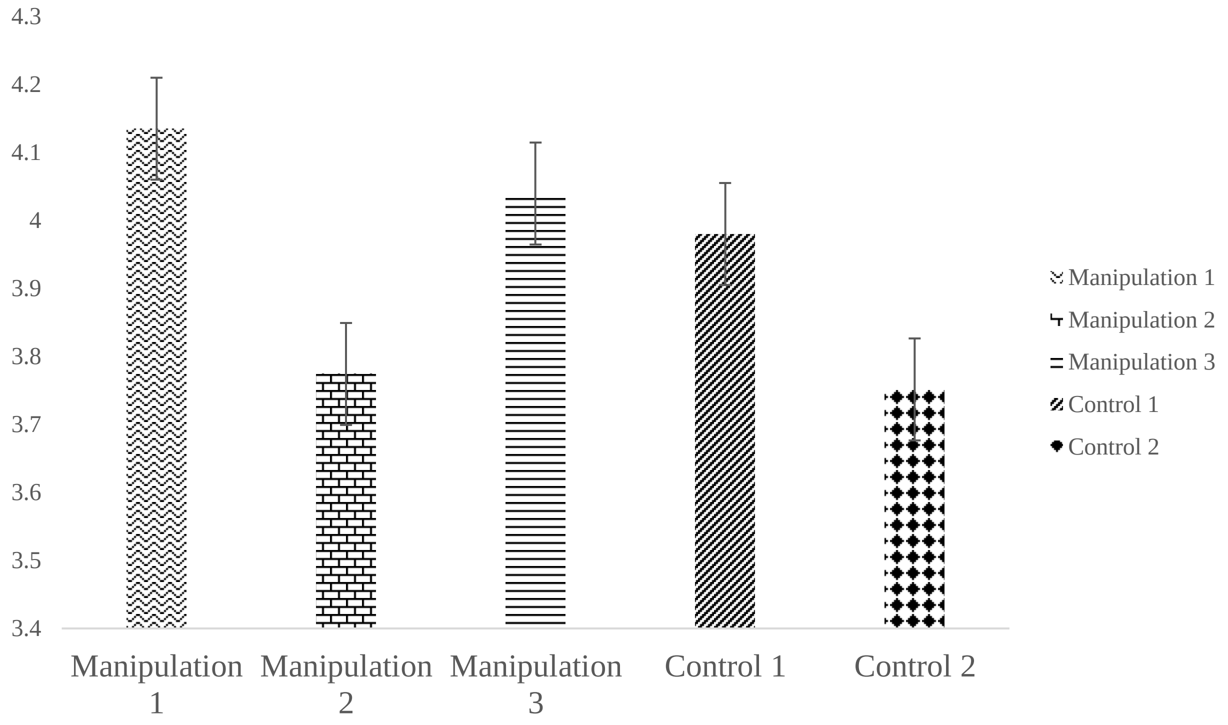


Figure 27. ANOVA Results of Manipulations on Willingness to Hire Minority Group Members in Study 2

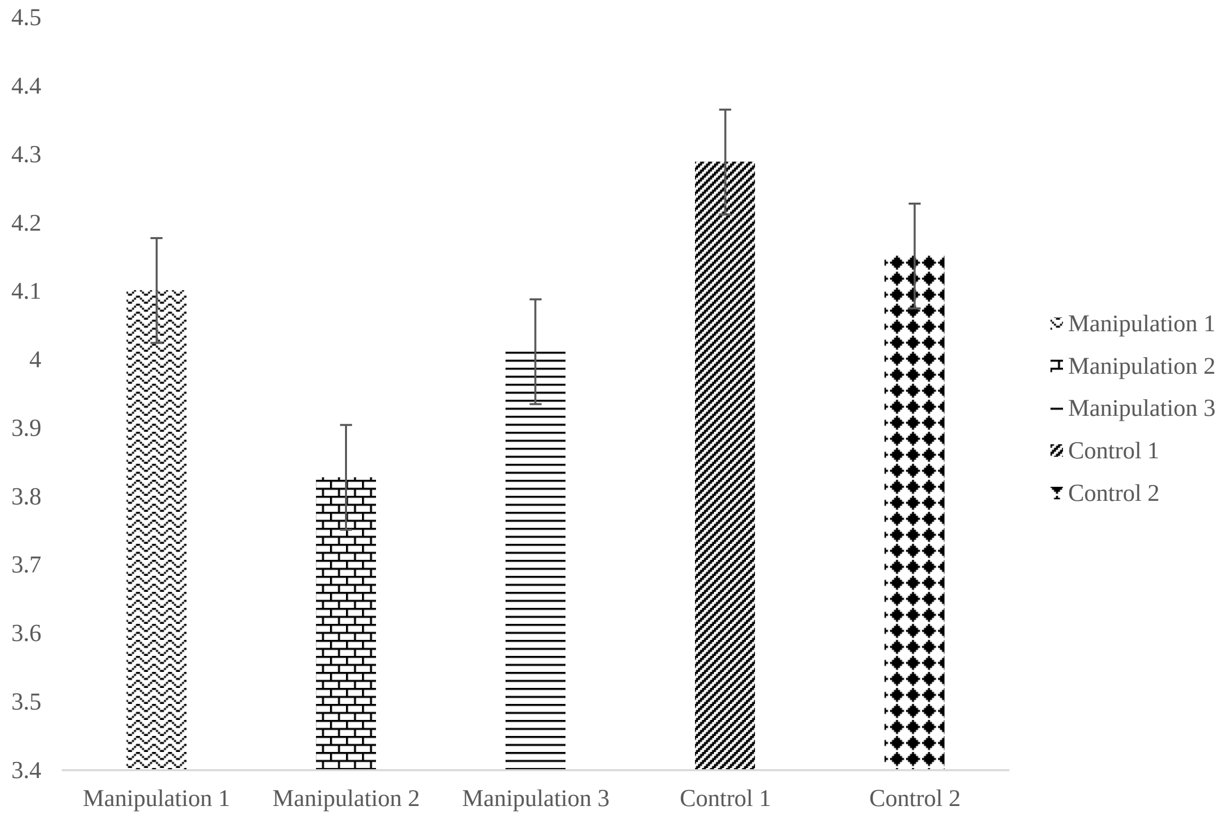


Table 14. Results of Direct Effects of Manipulations on DT Backlash Manifestations for Study 2

DVs	DT backlash manifestations															
	Post-IAT		MCD		MC		Entitle		Meta		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.45	1.59	0.22	0.14	-0.07	0.16	-0.09	0.10	0.38**	0.11	-0.19†	0.10	0.05	0.11	-0.05	0.17
Manipulation 2	-0.79	1.57	-0.07	0.14	-0.08	0.16	-0.04	0.10	0.02	0.11	-0.05	0.10	-0.09	0.10	-0.32†	0.17
Manipulation 3	1.34	1.61	0.35*	0.14	0.10	0.16	-0.03	0.10	0.29*	0.11	-0.21*	0.10	0.18†	0.11	-0.14	0.18
Control 1	-1.40	1.59	0.22†	0.14	-0.02	0.16	0.02	0.10	0.23*	0.11	-0.15	0.10	0.70	0.11	0.14	0.17
Pre-IAT	0.74**	0.04														
Intercept	-0.72	1.10	3.82**	0.09	2.94**	0.11	2.58**	0.07	3.76**		2.13**	0.07	4.07**	0.07	4.15**	0.12
R^2	0.45		0.03		0.01		0.01		0.04		0.01		0.01		0.02	

[a] *n* = 456. b = unstandardized regression coefficients. IVs were dummy variables, and the reference group was control group 2. Pretest-IAT was entered to test post-test IAT. MCD = moral credentials, MC = Moral credits, Entitle = psychological entitlement, Meta = metacognitive activity, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, Control 1 = control group 1.

* $p < .05$

** $p < .01$

† $p < .10$.

APPENDIX L STUDY 2 TABLES FOR MODERATION ANALYSES

Table 15. Results of Moderation Effects of Distributive Justice Perceptions between Manipulations and DT Backlash Manifestations for Study 2

DVs	DT backlash manifestations															
	Post-IAT		MCD		MC		Entitle		Meta		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.43	1.59	0.28^{**}	0.10	-0.04	0.15	-0.09	0.10	0.41^{**}	0.10	-0.21^{*}	0.09	0.07	0.09	-0.06	0.17
Manipulation 2	-0.92	1.57	0.15	0.10	0.09	0.15	-0.03	0.10	0.16	0.10	$-0.15^{\dagger}$	0.09	0.02	0.09	-0.34^{*}	0.17
Manipulation 3	1.11	1.61	0.31^{**}	0.10	0.05	0.15	-0.03	0.10	0.24^{*}	0.10	-0.19^{*}	0.09	0.13	0.10	-0.13	0.18
Control 1	-1.41	1.59	0.29^{**}	0.10	0.01	0.15	0.02	0.10	0.26^{**}	0.10	-0.18^{*}	0.09	0.10	0.09	0.13	0.17
DJ	-0.28	0.10	0.72^{**}	0.06	0.35^{**}	0.09	0.04	0.06	0.35^{**}	0.06	-0.33^{**}	0.06	0.31^{**}	0.06	-0.12	0.11
Pre-IAT	0.74^{**}	0.04														
Manipulation 1 x *DJ*	0.19	1.43	-0.30^{**}	0.10	-0.17	0.13	-0.07	0.09	$-0.15^{\dagger}$	0.09	0.06	0.08	-0.13	0.09	0.14	0.16
Manipulation 2 x *DJ*	-0.13	1.34	-0.13	0.08	0.15	0.12	-0.02	0.09	0.04	0.08	0.07	0.08	-0.01	0.08	0.08	0.15
Manipulation 3 x *DJ*	1.37	1.49	-0.20^{*}	0.09	0.05	0.14	0.04	0.09	0.03	0.09	0.11	0.08	0.04	0.09	0.02	0.16
Control 1 x *DJ*	3.34^{*}	1.43	-0.18^{*}	0.09	0.09	0.13	-0.08	0.09	-0.08	0.09	0.10	0.08	-0.01	0.09	0.25	0.16
Intercept	-0.70	1.10	3.76^{**}	0.07	2.91^{**}	0.10	2.58^{**}	0.07	3.72^{**}	0.07	2.16^{**}	0.06	4.05^{**}	0.07	4.16^{**}	0.12
R^2	0.46		0.50		0.18		0.01		0.27		0.21		0.22		0.02	

Table 15 (cont'd)

[a]*n* = 456. b = unstandardized regression coefficients. *SE* = standard error, MCD = moral credentials, MC = moral credits, Entitle = psychological entitlement, Meta = metacognitive activity, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, Control 1 = control group 1, DJ = distributive justice. IVs were dummy variables, and the reference group was control group 2. Pretest-IAT was entered to test post-test IAT.

* $p < .05$

** $p < .01$

† $p < .10$.

Table 16. Results of Moderation Effects of Procedural Justice Perceptions (assignment of DT) between Manipulations and DT Backlash Manifestations for Study 2

	DT backlash manifestations															
DVs	Post-IAT		MCD		MC		Entitle		Meta		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	B	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.48	1.59	0.16	0.12	-0.12	0.14	-0.10	0.10	0.35**	0.10	-0.17*	0.10	0.03	0.10	-0.05	0.17
Manipulation 2	-1.07	1.59	0.04	0.12	0.03	0.14	-0.02	0.10	0.10	0.10	-0.09†	0.10	-0.04	0.10	-0.31†	0.17
Manipulation 3	1.24	1.62	0.25*	0.12	-0.03	0.15	-0.05	0.10	0.19†	0.11	-0.17*	0.10	0.11	0.10	-0.13	0.18
Control 1	-1.45	1.59	0.19	0.12	-0.06	0.14	0.02	0.10	0.21*	0.10	-0.14*	0.10	0.05	0.10	0.14	0.17
PJa	-0.18	1.15	0.55**	0.09	0.56**	0.10	0.04	0.07	0.34**	0.07	-0.17*	0.07	0.24**	0.07	-0.16	0.13
Pre-IAT	0.74**	0.04														
Manipulation 1 x *PJa*	-0.03	1.75	-0.11	0.13	-0.12	0.16	0.05	0.11	-0.05†	0.11	0.02	0.11	-0.09	0.11	0.19	0.19
Manipulation 2 x *PJa*	-1.06	1.64	0.03	0.12	0.01	0.15	0.04	0.10	0.05	0.11	-0.01	0.10	0.01	0.10	0.20	0.18
Manipulation 3 x *PJa*	0.73	1.75	-0.18	0.13	-0.04	0.16	0.06	0.11	0.08	0.11	0.02	0.11	0.01	0.11	0.14	0.19
Control 1 x *PJa*	3.52*	1.72	-0.07	0.13	0.03	0.15	-0.15	0.11	-0.01	0.11	-0.02	0.11	0.14	0.11	0.36†	0.19
Intercept	-0.75	1.10	3.84**	0.08	2.97**	0.10	2.58**	0.07	3.77**	0.07	2.13**	0.07	4.08**	0.07	4.15**	0.12
R^2	0.46		0.26		0.21		0.02		0.21		0.06		0.13		0.03	

Table 16 (cont'd)

[a]*n* = 456. b = unstandardized regression coefficients. *SE* = standard error, MCD = moral credentials, MC = moral credits, Entitle = psychological entitlement, Meta = metacognitive activity, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, Control 1 = control group 1, PJa = procedural justice regarding the assignment of DT. IVs were dummy variables, and the reference group was control group 2. Pretest-IAT was entered to test post-test IAT.

[*] $p < .05$

[**] $p < .01$

[†] $p < .10$.

Table 17. Results of Moderation Effects of SDO between Manipulations and DT Backlash Manifestations for Study 2

	DT backlash manifestations															
DVs	Post-IAT		MCD		MC		Entitle		Meta		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.36	1.58	0.11	0.11	-0.08	0.16	-0.06	0.10	0.36^{**}	0.11	-0.08	0.07	-0.02	0.09	-0.04	0.18
Manipulation 2	-0.71	1.56	-0.01	0.11	-0.04	0.16	-0.06	0.10	0.06	0.11	-0.08	0.07	-0.03	0.09	-0.33	0.18
Manipulation 3	1.25	1.58	0.29^{*}	0.11	0.08	0.16	-0.02	0.10	0.26^{*}	0.11	-0.15^{*}	0.07	0.14	0.09	-0.12	0.18
Control 1	-1.58	1.56	0.17	0.11	-0.03	0.16	0.03	0.10	$0.21^{\dagger}$	0.11	-0.10	0.07	0.03	0.09	0.15	0.17
SDO	-0.59	1.48	-0.91^{**}	0.11	-0.11	0.15	0.19	0.09	-0.40^{**}	0.10	0.83^{**}	0.07	-0.59^{**}	0.09	$0.27^{\dagger}$	0.16
Pre-IAT	0.74^{**}	0.04														
Manipulation 1 x *SDO*	-0.12	2.45	0.24	0.17	0.03	0.24	0.03	0.15	$0.30^{\dagger}$	0.17	-0.16	0.12	0.19	0.14	-0.29	0.27
Manipulation 2 x *SDO*	-0.36	2.21	-0.05	0.16	-0.32	0.22	0.14	0.14	-0.15	0.15	-0.25^{*}	0.10	-0.18	0.14	-0.07	0.25
Manipulation 3 x *SDO*	-1.12	2.30	0.27	0.16	-0.31	0.23	-0.08	0.14	0.08	0.16	-0.25^{*}	0.11	0.27^{*}	0.13	-0.14	0.26
Control 1 x *SDO*	-7.48^{**}	2.33	0.09	0.17	$-0.38^{\dagger}$	0.23	0.13	0.15	0.12	0.16	-0.16	0.11	-0.02	0.13	-0.10	0.26
Intercept	-0.71	1.08	3.85^{**}	0.08	2.95^{**}	0.11	2.58^{**}	0.07	3.77^{**}	0.07	2.10^{**}	0.05	4.10^{**}	0.06	4.15^{**}	0.12
R^2	0.47		0.36		0.05		0.06		0.14		0.46		0.29		0.03	

Table 17 (cont'd)

[a]*n* = 456. b = unstandardized regression coefficients. *SE* = standard error, MCD = moral credentials, MC = moral credits, Entitle = psychological entitlement, Meta = metacognitive activity, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, Control 1 = control group 1, SDO = social dominance orientation. IVs were dummy variables, and the reference group was control group 2. Pretest-IAT was entered to test post-test IAT.

* $p < .05$

** $p < .01$

† $p < .10$.

Table 18. Results of Moderation Effects of BJW between Manipulations and DT Backlash Manifestations for Study 2

	DT backlash manifestations															
DVs	Post-IAT		MCD		MC		Entitle		Meta		Prejudice		OCB		Hire	
IVs	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*	b	*SE*
Manipulation 1	0.45	1.60	0.25	0.13	-0.05	0.16	-0.10	0.10	0.39**	0.11	-0.19*	0.09	0.06	0.10	-0.04	0.17
Manipulation 2	-0.73	1.58	-0.04	0.13	-0.10	0.15	-0.05	0.10	0.03	0.11	-0.09	0.09	-0.07	0.10	-0.32†	0.17
Manipulation 3	1.44	1.61	0.39*	0.14	0.08	0.16	-0.04	0.10	0.29*	0.11	-0.25**	0.09	0.19†	0.11	-0.14	0.18
Control 1	-1.02	1.61	0.26†	0.13	-0.07	0.16	-0.02	0.10	0.20†	0.11	-0.21*	0.09	0.10	0.10	0.10	0.17
BJW	-0.79	1.73	-0.44**	0.15	0.24	0.17	0.13	0.11	-0.04	0.12	0.50**	0.10	-0.24*	0.11	0.03	0.19
Pre-IAT	0.74**	0.04														
Manipulation 1 x *BJW*	0.01	2.47	0.47*	0.21	0.30	0.24	-0.09	0.15	0.12†	0.18	-0.10	0.14	0.16	0.16	0.12	0.27
Manipulation 2 x *BJW*	-0.38	2.45	0.13	0.21	0.24	0.24	0.28†	0.15	-0.02	0.17	-0.12	0.14	-0.16	0.16	0.47†	0.27
Manipulation 3 x *BJW*	-0.77	2.58	-0.01	0.22	-0.14	0.25	-0.07	0.16	0.06	0.18	-0.22	0.15	0.11	0.17	0.15	0.28
Control 1 x *BJW*	-2.40	2.42	0.35†	0.20	0.13	0.24	0.15	0.15	0.33†	0.17	-0.22	0.14	0.11	0.16	0.41	0.26
Intercept	-0.78	1.11	3.85**	0.08	2.96**	0.11	2.58**	0.07	3.75**	0.08	2.16**	0.06	4.05**	0.07	4.15**	0.12

Table 18 (cont'd)

	Post-IAT	MCD	MC	Entitle	Meta	Prejudice	OCB	Hire
R^2	0.45	0.07	0.06	0.05	0.15	0.46	0.06	0.03

[a]*n* = 456. b = unstandardized regression coefficients. *SE* = standard error, MCD = moral credentials, MC = moral credits, Entitle = psychological entitlement, Meta = metacognitive activity, Prejudice = explicit prejudice, OCB = organizational citizenship behavior, Hire = willingness to hire minority group members, Control 1 = control group 1, BJW = belief in a just world. IVs were dummy variables, and the reference group was control group 2. Pretest-IAT was entered to test post-test IAT.

* $p < .05$

** $p < .01$

† $p < .10$.

APPENDIX M STUDY 2 FIGURES FOR MODERATION ANALYSES

Figure 28. Interaction between Manipulation Group 1 and Distributive Justice on Moral Credentials (Study 2)

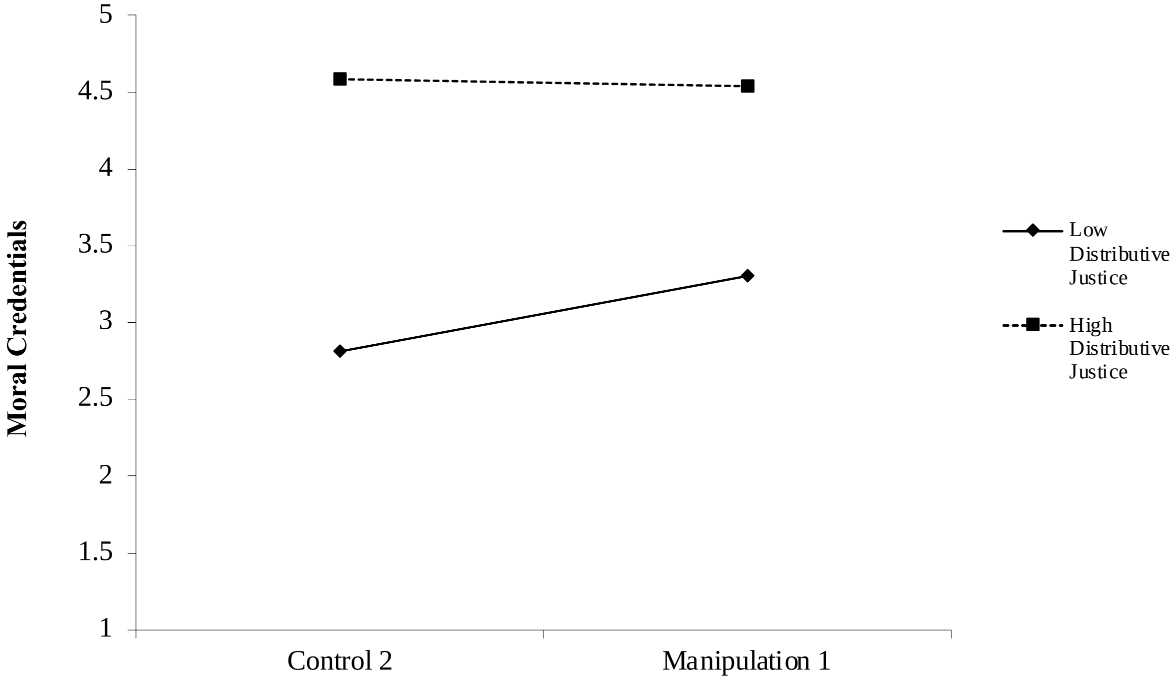


Figure 29. Interaction between Manipulation Group 3 and Distributive Justice on Moral Credentials (Study 2)

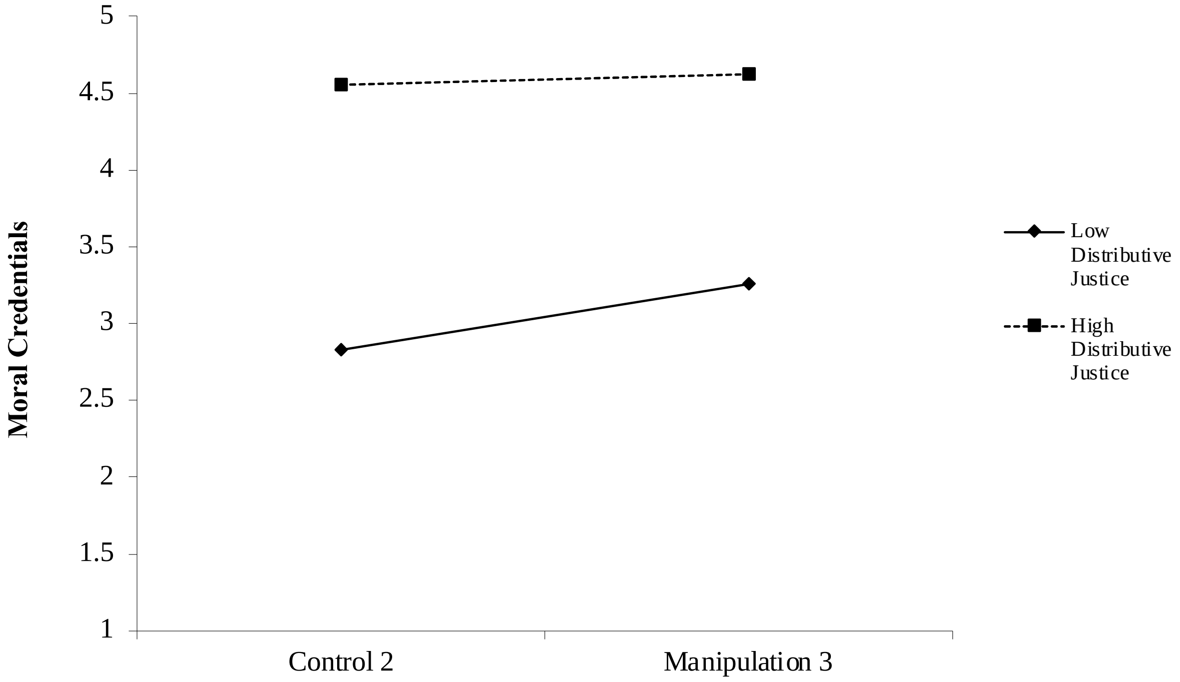


Figure 30. Interaction between Control Group 1 and Distributive Justice on Moral Credentials (Study 2)

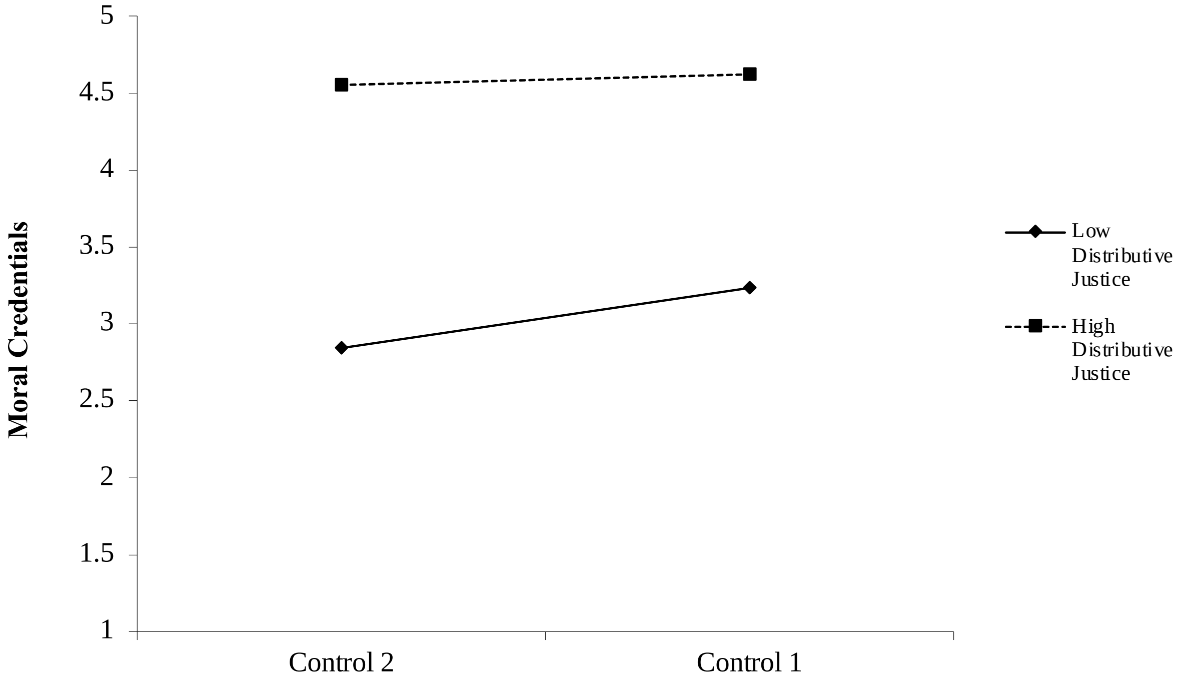


Figure 31. Interaction between Manipulation Group 1 and Distributive Justice on Anger (Study 2)

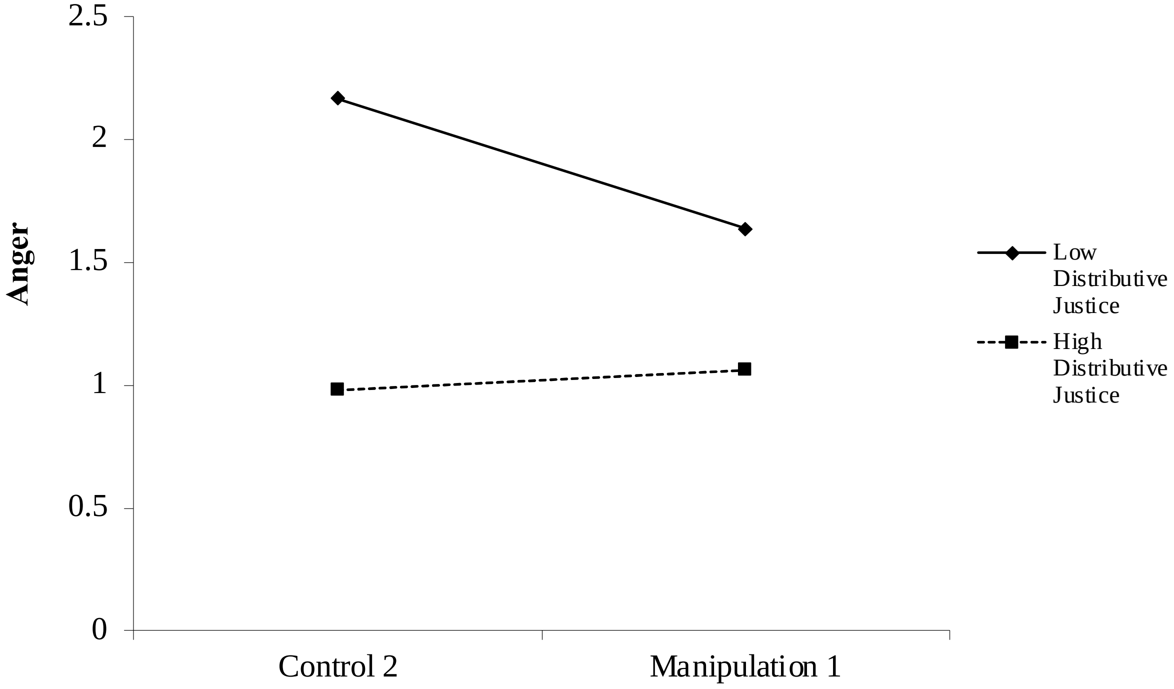


Figure 32. Interaction between Manipulation Group 3 and Distributive Justice on Anger (Study 2)

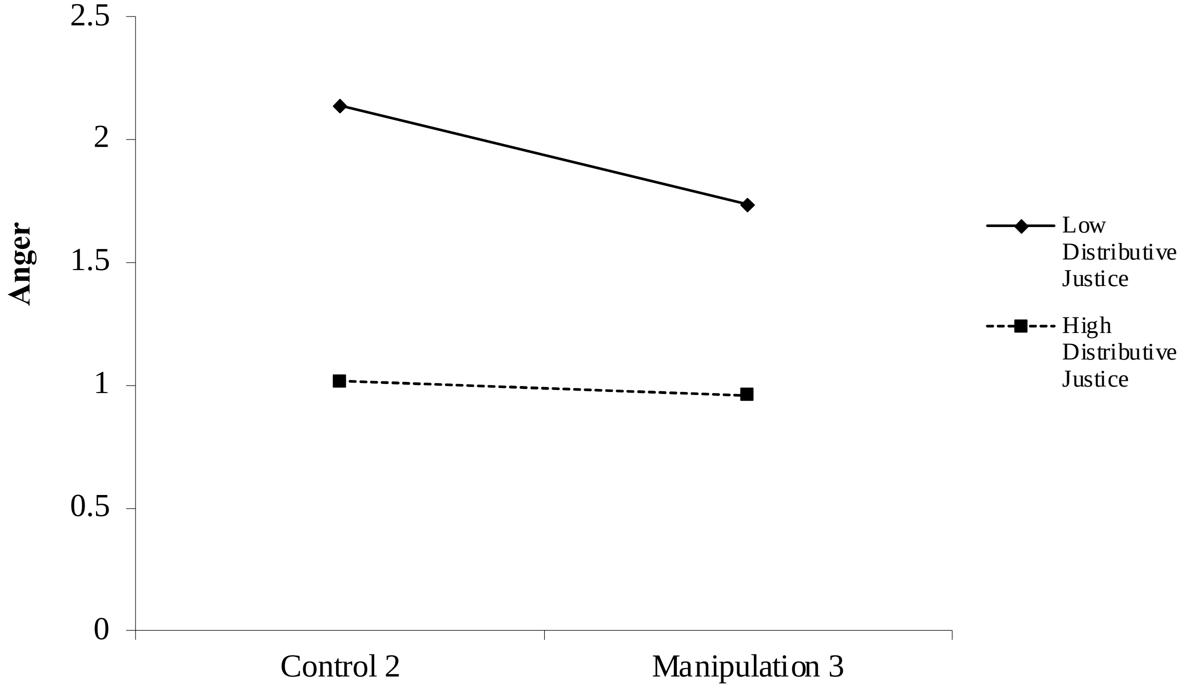


Figure 33. Interaction between Control Group 1 and Distributive Justice on Anger (Study 2)

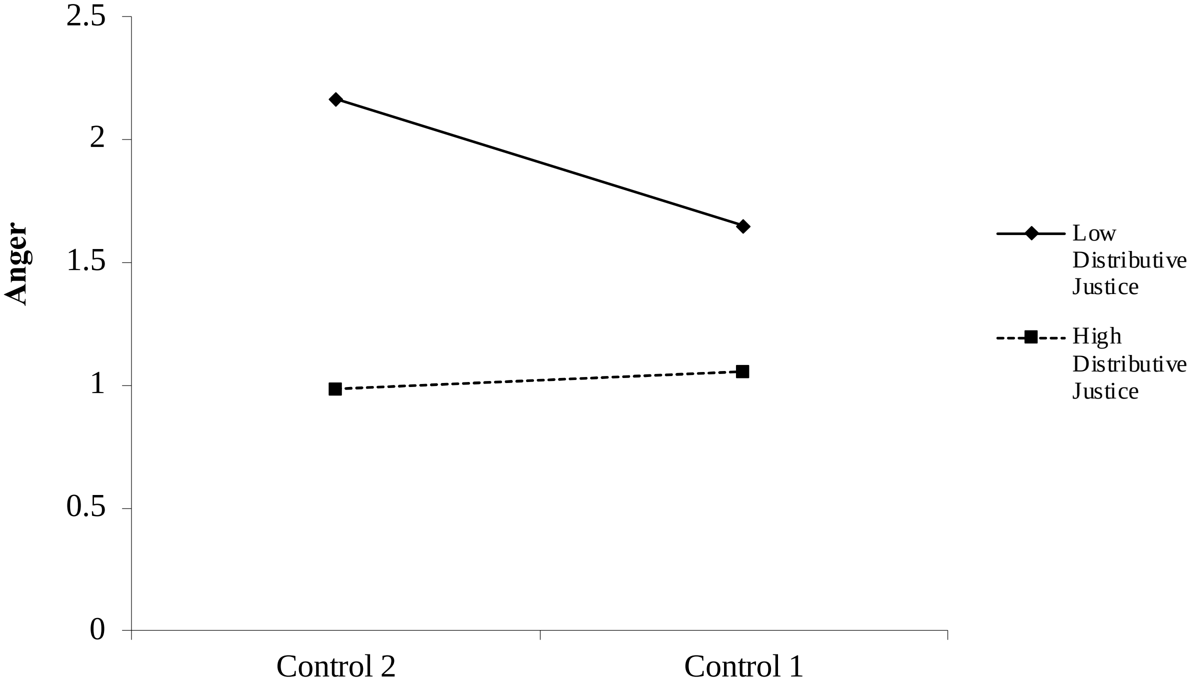


Figure 34. Interaction between Manipulation Group 1 and Distributive Justice on Metacognitive Activity (Study 2)

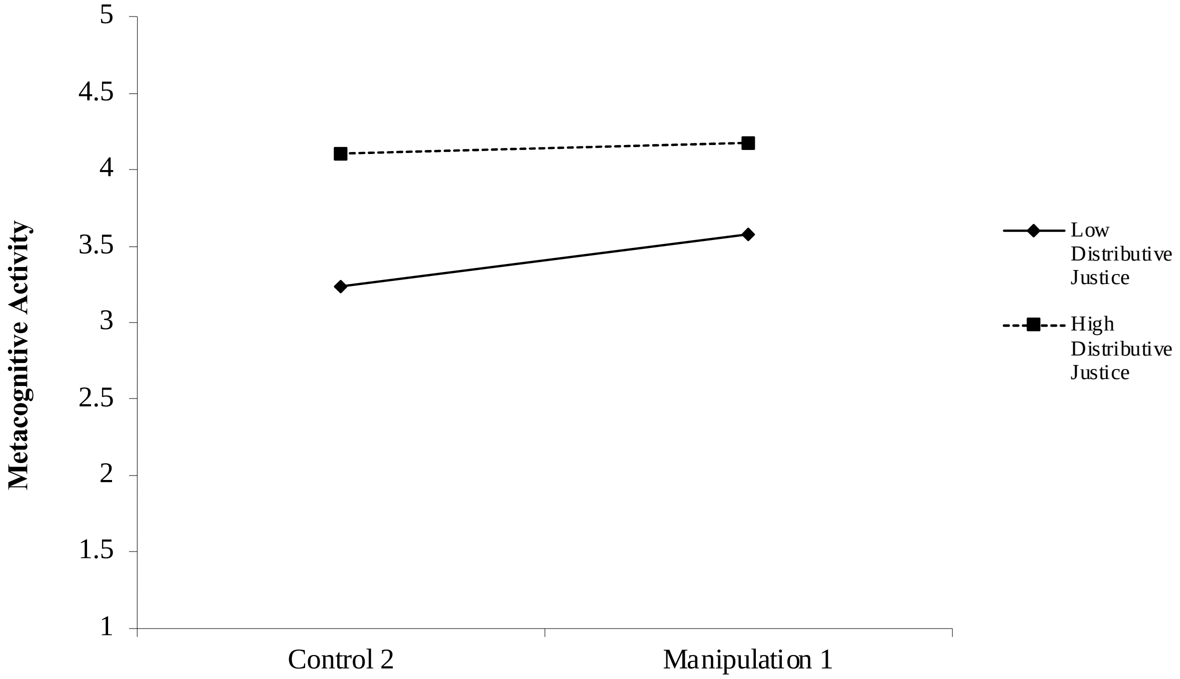


Figure 35. Interaction between Manipulation Group 1 and Procedural Justice on Anger (Study 2)

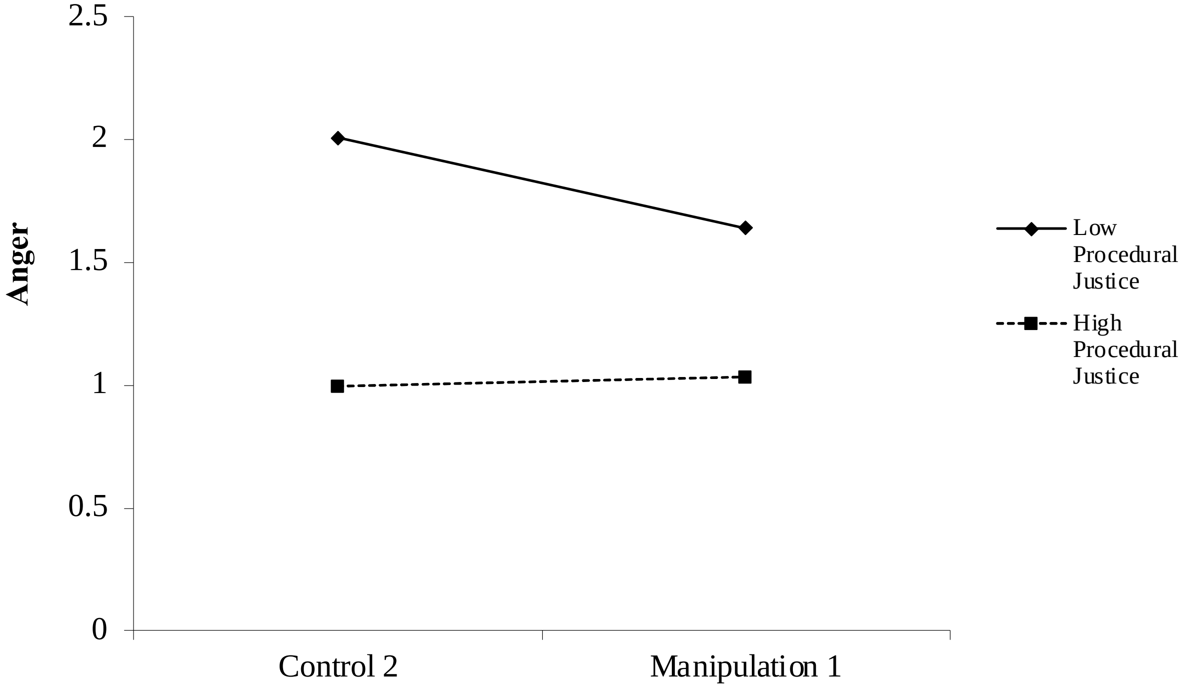


Figure 36. Interaction between Manipulation Group 3 and Procedural Justice on Anger (Study 2)

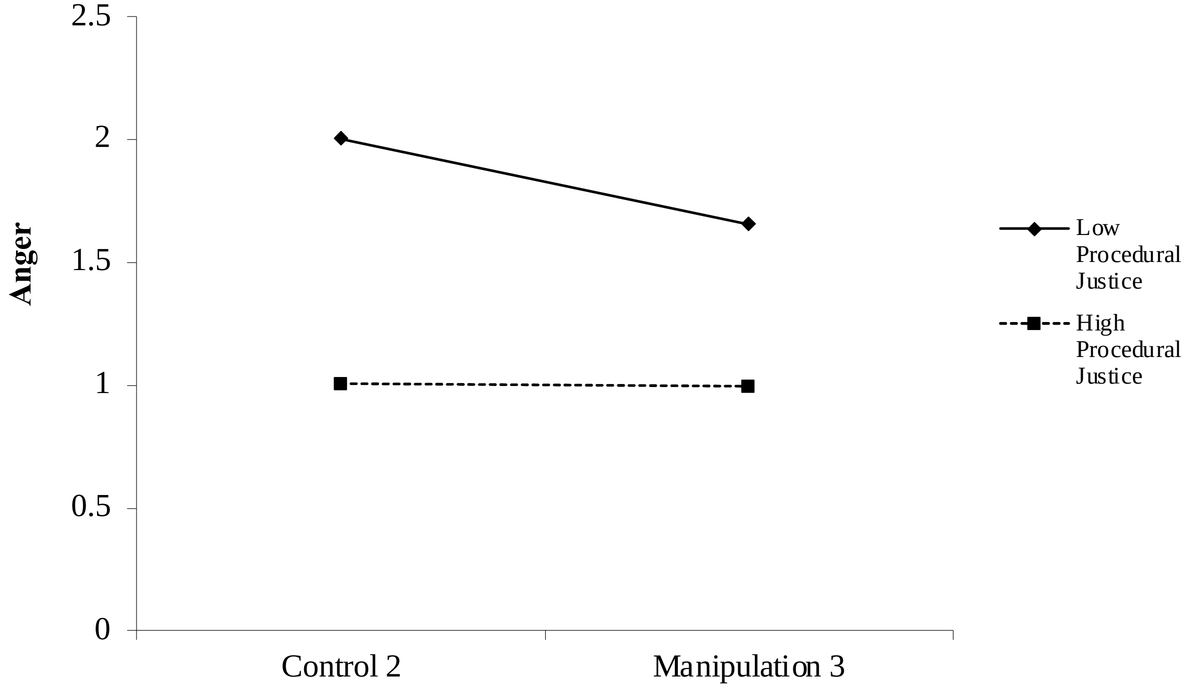


Figure 37. Interaction between Control Group 1 and Procedural Justice on Anger (Study 2)

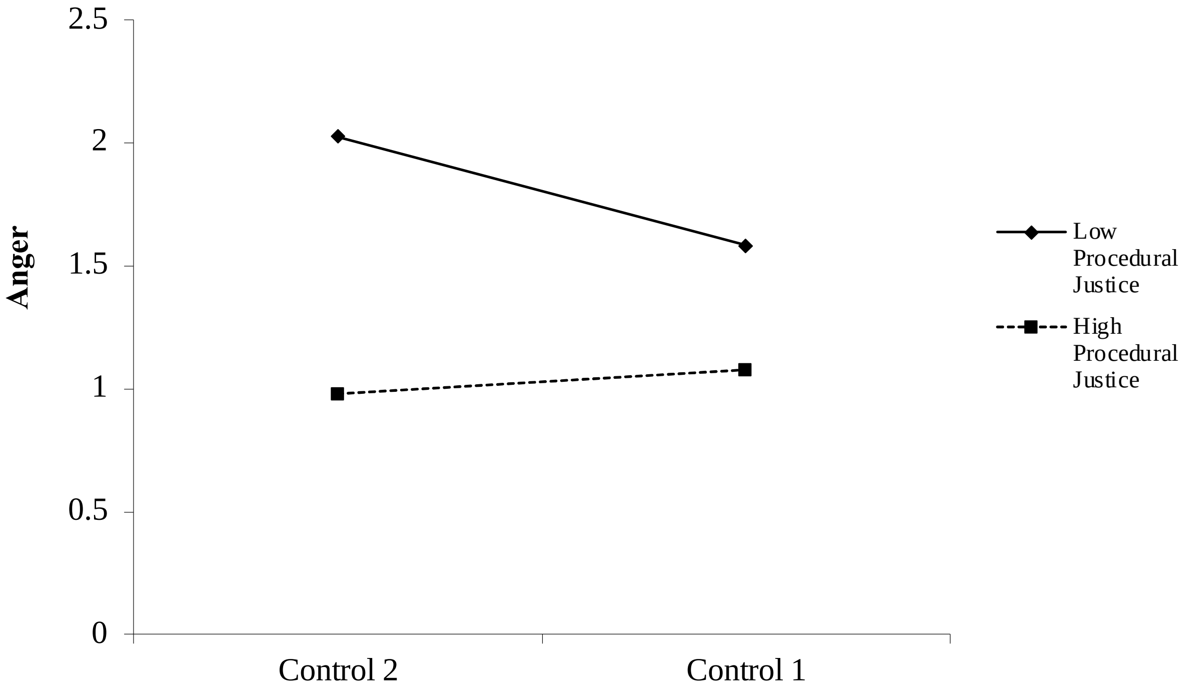


Figure 38. Interaction between Control Group 1 and Procedural Justice on IAT (Study 2)

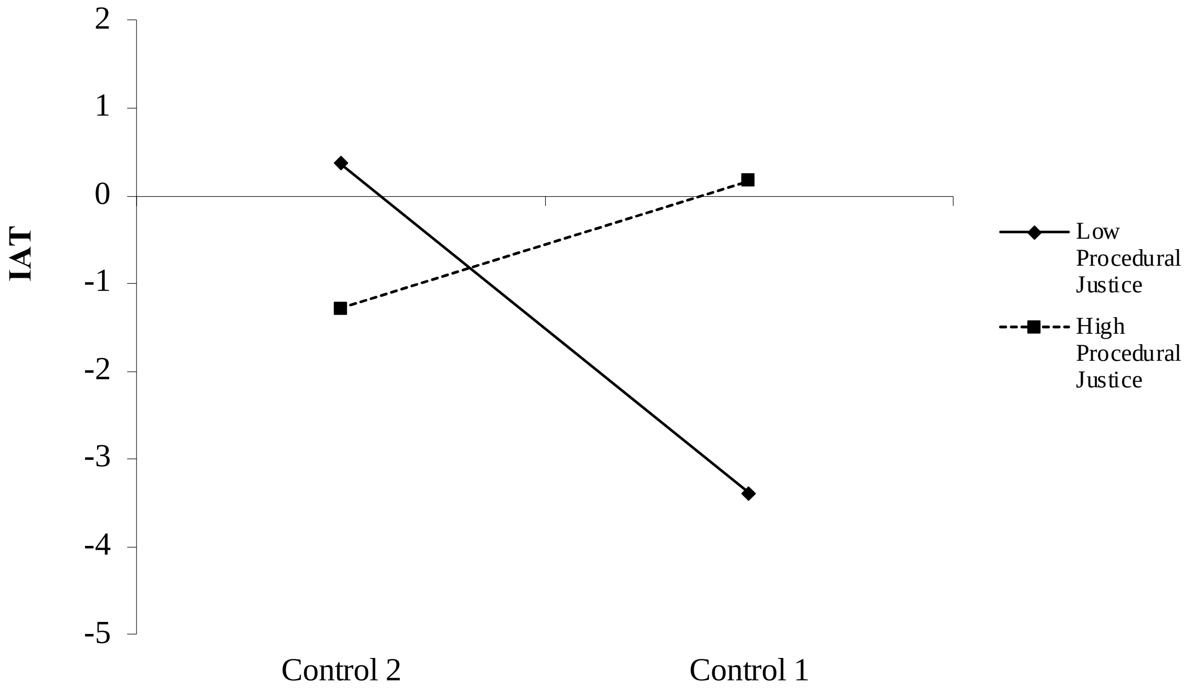


Figure 39. Interaction between Manipulation Group 1 and Social Dominance Orientation on Anger (Study 2)

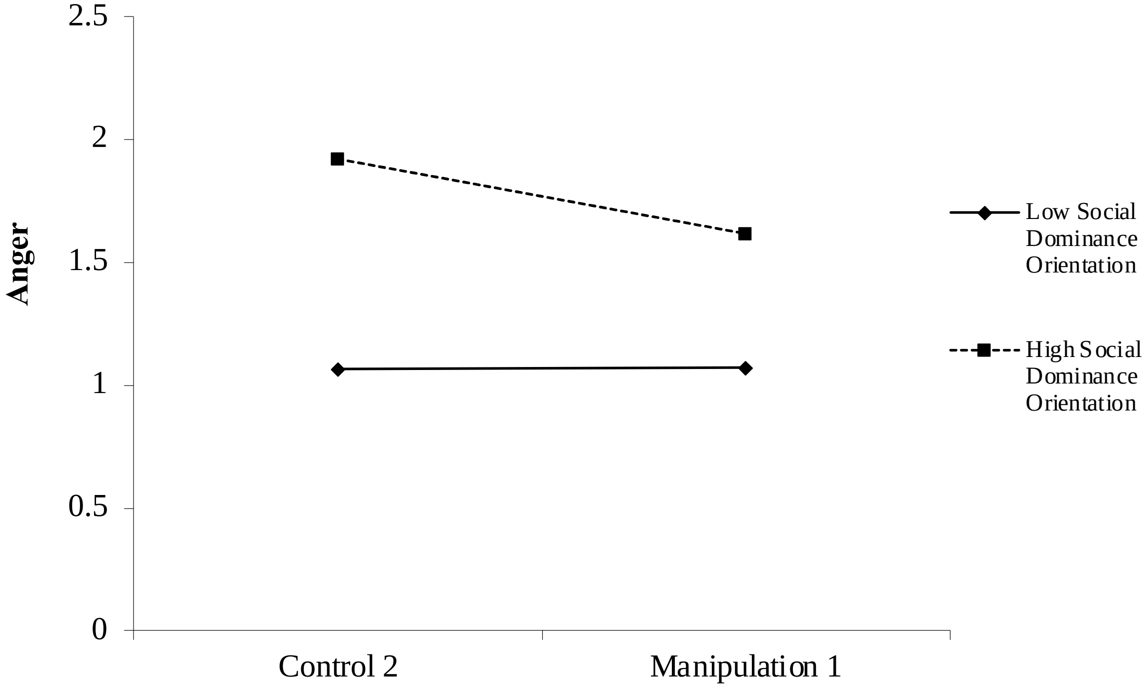


Figure 40. Interaction between Manipulation Group 2 and Social Dominance Orientation on Explicit Prejudice (Study 2)

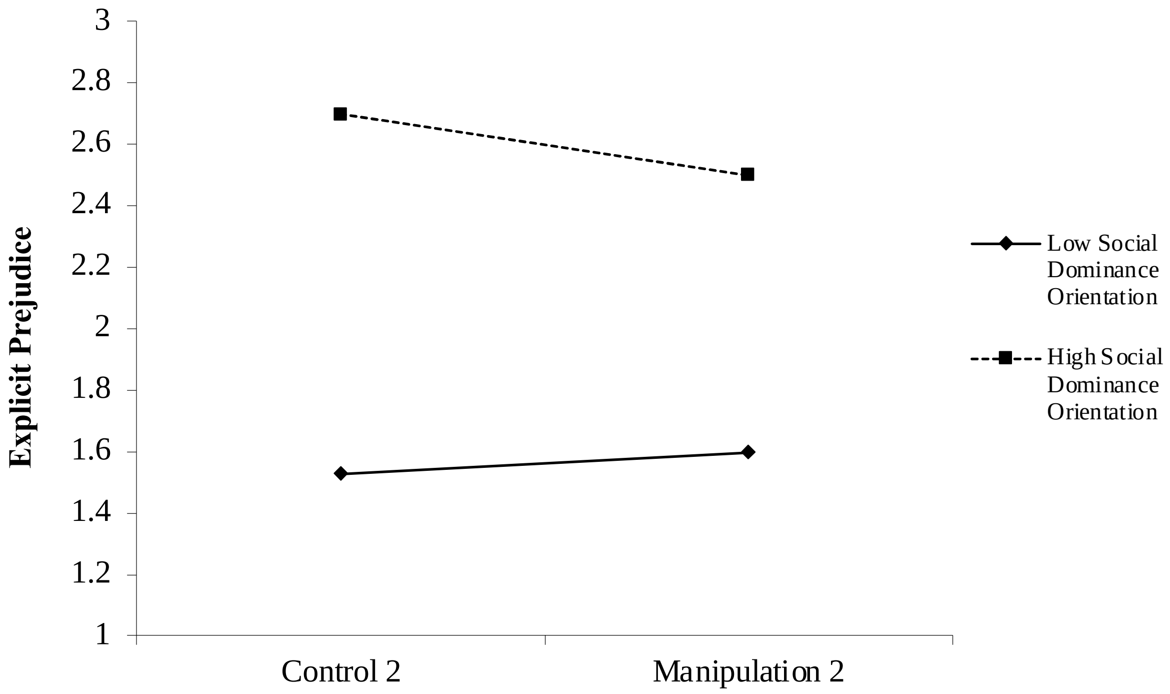


Figure 41. Interaction between Manipulation Group 3 and Social Dominance Orientation on Explicit Prejudice (Study 2)

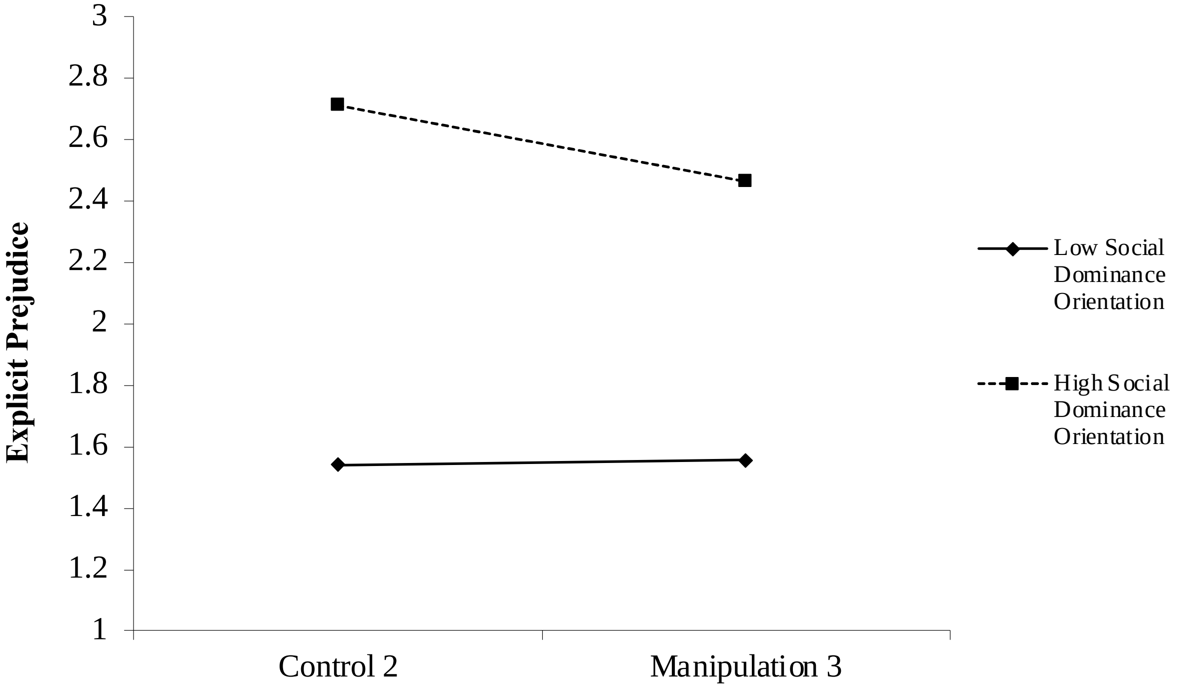


Figure 42. Interaction between Manipulation Group 3 and Social Dominance Orientation on OCB (Study 2)

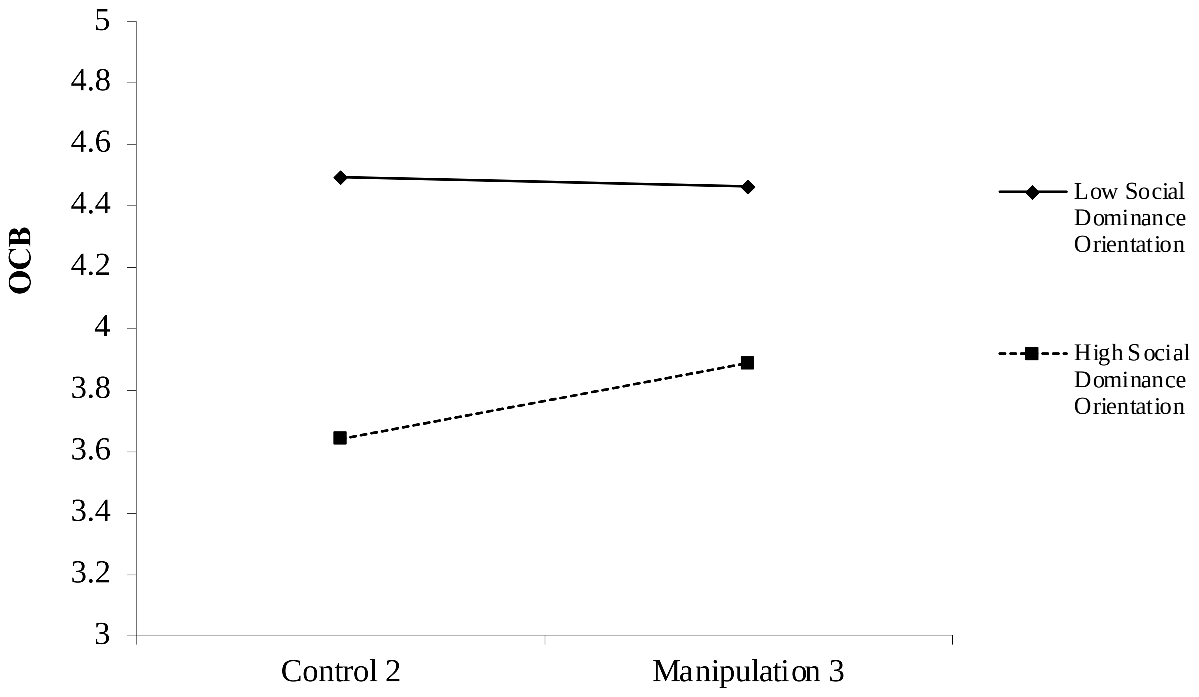


Figure 43. Interaction between Manipulation Group 1 and Social Dominance Orientation on Metacognitive Activity (Study 2)

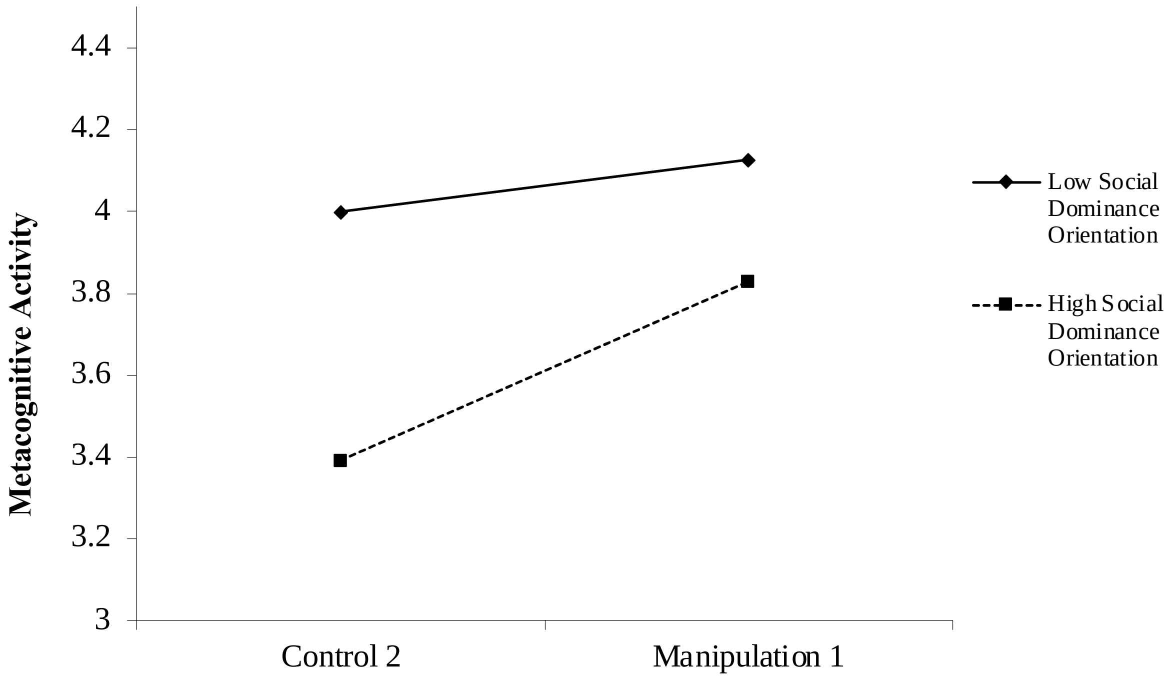


Figure 44. Interaction between Control Group 1 and Social Dominance Orientation on Implicit Prejudice (IAT score; Study 2)

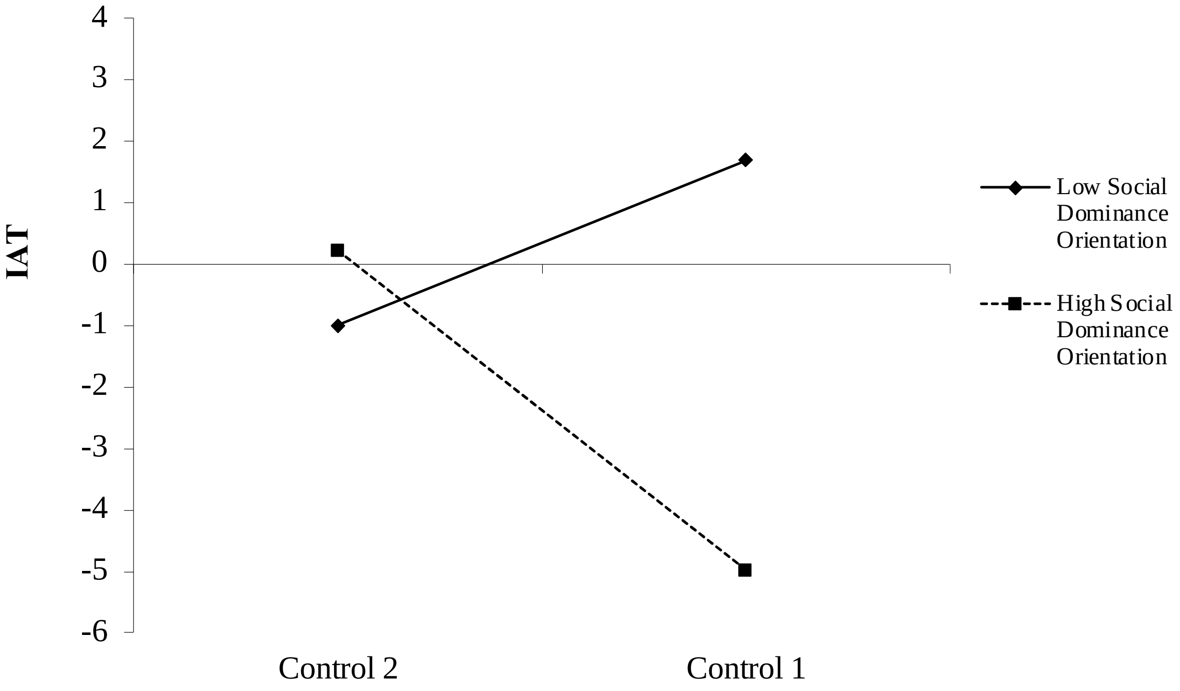


Figure 45. Interaction between Manipulation Group 1 and Belief in a Just World on Moral Credentials (Study 2)

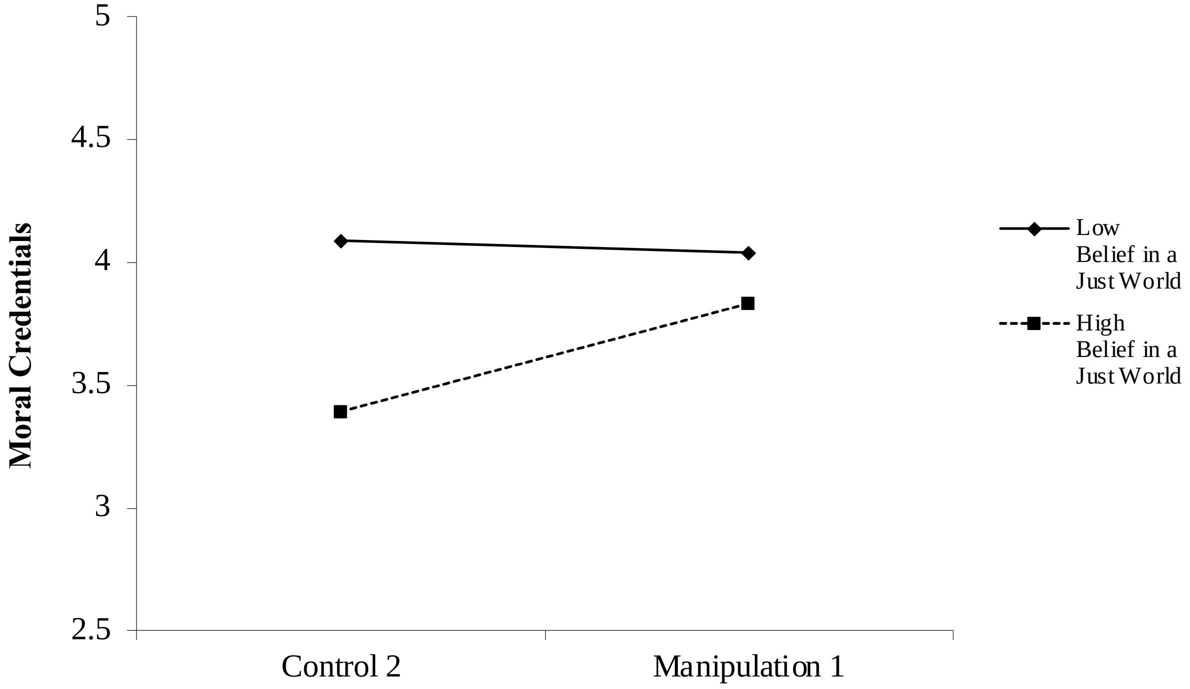


Figure 46. Interaction between Control Group 1 and Belief in a Just World on Moral Credentials (Study 2)

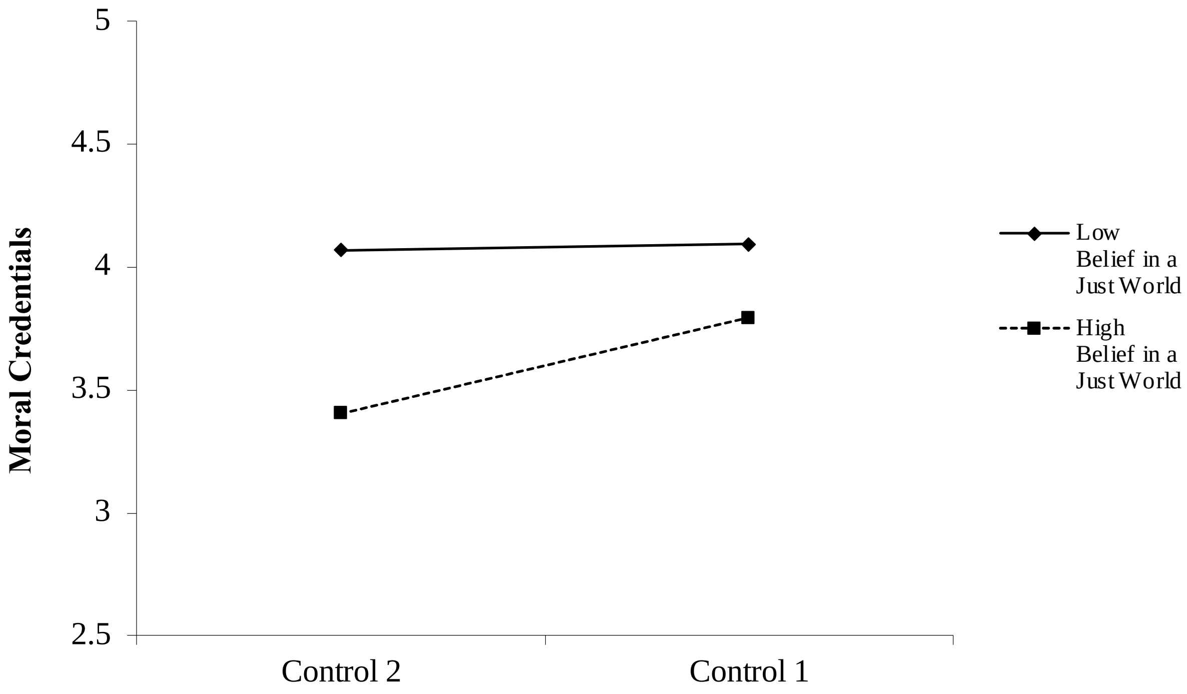


Figure 47. Interaction between Manipulation Group 2 and Belief in a Just World on Psychological Entitlement (Study 2)

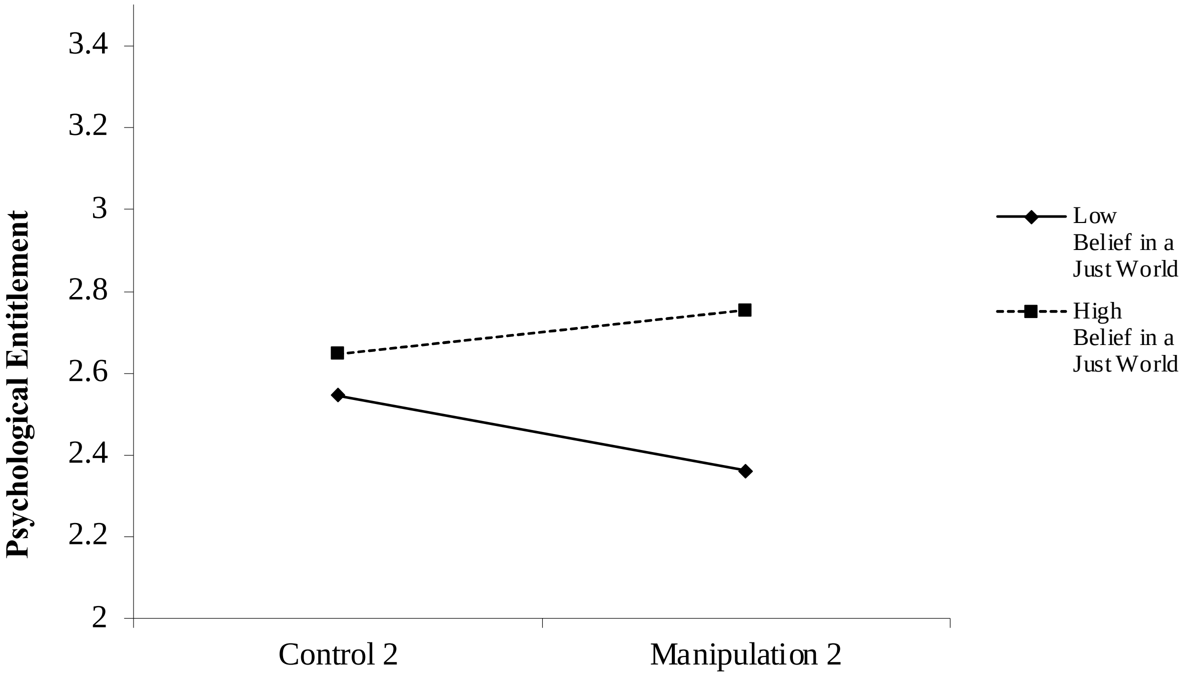
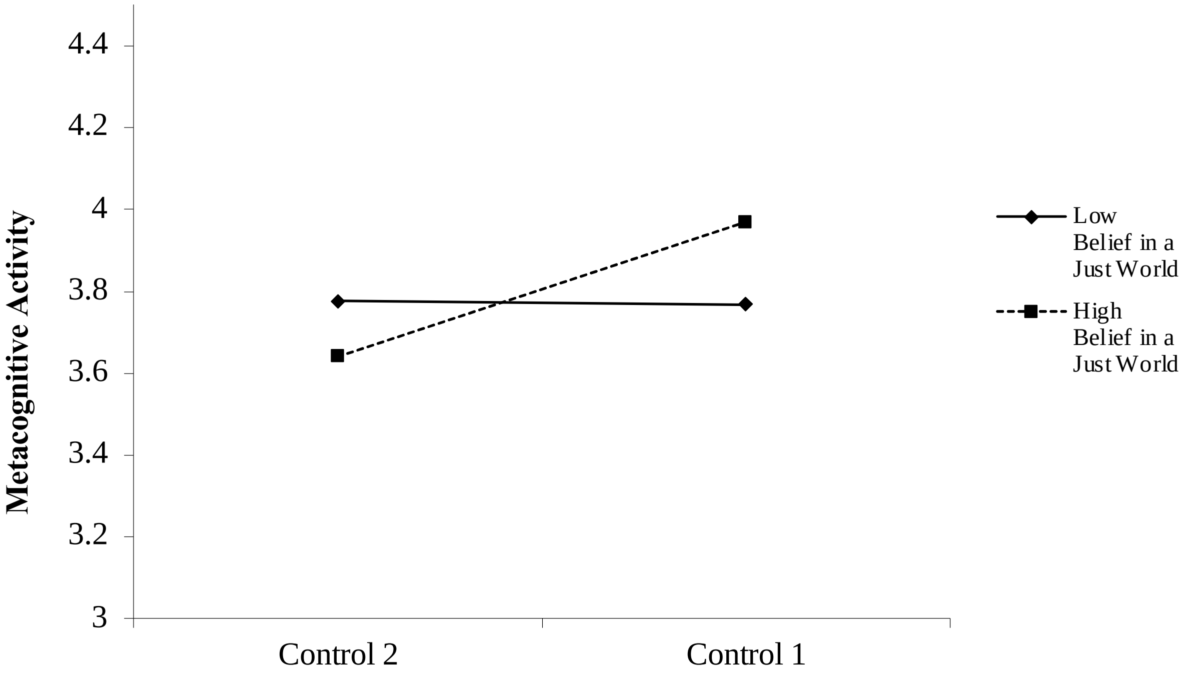


Figure 48. Interaction between Control Group 1 and Belief in a Just World on Metacognitive Activity (Study 2)



APPENDIX N SUMMARY OF RESULTS FOR STUDY 1 AND 2

Table 19. Summary of Direct Effects and Moderating Effects on DT Backlash in Study 1

DVs	DT Backlash							
	Implicit prejudice	Moral credentials	Moral credits	Entitlement	Anger	Explicit prejudice	OCB	Hiring decision
Mani 1								
Mani 2						Consistency		
Mani 3					Consistency	Consistency		Licensing
Control								
Mani 1 x DJ		Low DJ increased moral credentials; high DJ ns.			Low DJ decreased anger; high DJ ns.			
Mani 2 x DJ		Low DJ increased moral credentials; high DJ ns.				Low DJ decreased prejudice; high DJ ns.		
Mani 3 x DJ		Low DJ increased moral credentials; high DJ ns.				Low DJ decreased prejudice; high DJ ns.		
Mani 1 x PJa								
Mani 2 x PJa								
Mani 3 x PJa		Low PJa increased moral credentials; high PJa ns.						

Table 19 (cont'd)

DVs	Implicit prejudice	Moral credentials	Moral credits	Entitlement	Anger	Explicit prejudice	OCB	Hiring decision
Mani 1 x PJd								
Mani 2 x PJd						Low PJd decreased prejudice; high DJ ns		
Mani 3 x PJd		Low PJd increased moral credentials; high PJd ns.				Low PJd decreased prejudice; high DJ ns	High PJd decreased OCB (p <. 10); low PJd ns.	
Mani 1 x SDO								
Mani 2 x SDO	High SDO decreased implicit prejudice; low SDO ns.			High SDO increased entitlement; low SDO ns.				
Mani 3 x SDO				High SDO increased entitlement; low SDO ns.				
Mani 1 x BJW						High BJW decreased explicit prejudice; low BJW ns.		
Mani 2 x BJW		High BJW increased moral credentials; low SDO ns.				High BJW decreased explicit prejudice; low BJW ns.		
Mani 3 x BJW				Both simple slopes ns.		High BJW decreased explicit prejudice; low BJW ns.		

Table 20. Summary of Direct Effects and Moderating Effects on DT Backlash in Study 2

	DT Backlash								
DVs	Implicit prejudice	Moral credentials	Moral credits	Psychological entitlement	Anger	Explicit prejudice	OCB	Hiring decision	Metacognitive activity
Mani 1					Consistency	Consistency			Consistency
Mani 2								Consistency	
Mani 3		Consistency			Consistency	Consistency	Consistency		Consistency
Control 1		Consistency			Consistency				Consistency
Control 2									
Mani 1 x DJ		Low DJ increased moral credentials; high DJ ns.			Low DJ decreased anger; high DJ ns.				Low DJ increased metacognitive activity; high DJ ns.
Mani 2 x DJ									
Mani 3 x DJ		Low DJ increased moral credentials; high DJ ns.			Low DJ decreased anger; high DJ ns.				
Control 1 x DJ		Low DJ increased moral credentials; high DJ ns.			Low DJ decreased anger; high DJ ns.				
Mani 1 x PJa					Low PJ decreased anger; high PJ ns.				
Mani 2 x PJa									
Mani 3 x PJa					Low PJ decreased anger; high PJ ns.				

Table 20 (cont'd)

DVs	Implicit prejudice	Moral credentials	Moral credits	Psychological entitlement	Anger	Explicit prejudice	OCB	Hiring decision	Metacognitive activity
Control 1 x PJa	Both slopes decreased implicit prejudice, but low PJ decreased more				Low PJ decreased anger; high PJ ns.				
Mani 1 x SDO					High SDO decreased anger; low SDO ns.				High SDO increased metacognitive activity; low SDO ns.
Mani 2 x SDO						High SDO decreased explicit prejudice; low SDO ns.			
Mani 3 x SDO						High SDO decreased explicit prejudice; low SDO ns	High SDO increased OCB; low SDO ns.		
Control 1 x SDO	High SDO decreased implicit prejudice (p < .10); low SDO increased prejudice		Both slopes ns.						
Mani 1 x BJW		High BJW increased moral credentials; low BJW ns.							

Table 20 (cont'd)

DVs	Implicit prejudice	Moral credentials	Moral credits	Psychological entitlement	Anger	Explicit prejudice	OCB	Hiring decision	Metacognitive activity
Mani 2 x BJW				Low BJW decreased entitlement (p < .10); high BJW ns.					
Mani 3 x BJW									
Control 1 x BJW		High BJW increased moral credentials; low BJW ns.							High BJW increased metacognitive activity; low BJW ns.