WASHINGTON UNIVERSITY IN SAINTLOUIS

School of Arts & Sciences
Department of Political Science

Dissertation Examination Committee:
Jacob Montgomery, Chair
Taylor Carlson
Betsy Sinclair
Morgan Hazelton
Ted Enamorado

Cognitive Landscapes:
Argument Evaluations, Misinformation Corrections, and Racial Attitudes in Modern Media

by

Dominique Lockett

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

July 2024
Saint Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation owes its shape and substance to the guidance and support of several key individuals and the entire Political Science department at Washington University in Saint Louis. Their collective expertise and dedication have profoundly enriched my academic journey.

I am particularly grateful to Jacob Montgomery, whose eagerness to teach was evident from the outset. Jacob set high expectations and provided a diverse array of projects that significantly broadened my technical expertise. His passion for academia and persistence in advocating for rigorous mathematical training have been instrumental in my development as a scholar. Despite my initial lack of mathematical background, Jacob's commitment to teaching through numerous encounters helped build my intuition and understanding, embodying the true spirit of a great educator. His meticulous attention to refining my skills in presentation and analysis has left an indelible mark on both my academic and professional persona.

I extend my deepest gratitude to Morgan Hazelton, whose encouragement to pursue advanced education profoundly influenced my decision to attend Washington University in Saint Louis. Throughout my time as an undergraduate at Saint Louis University, Morgan served as a steadying influence and her insightful guidance was pivotal in steering me toward this rewarding path. As I went on to pursue my doctoral journey, I discovered a true excitement for data science and problem-solving through logical approaches. I could not have discovered this personal passion without Morgan's gentle and accessible method of teaching political

methodologies. Her influence in sparking my interest in computational science has been instrumental in shaping, not only, my academic and professional trajectory, but also pursuits of personal passion, for which I am immensely grateful.

I extend heartfelt thanks to Betsy Sinclair, whose deep empathy provided crucial support during the most challenging periods of my journey, so far. Her unwavering support served as a guiding light, helping me navigate both personal and academic challenges with resilience and reminding me of the profound impact of compassionate mentorship.

I am thankful to Taylor Carlson for consistently providing constructive and timely feedback on my work. Taylor's detailed insights were crucial in refining my arguments and improving the clarity of my research, thereby enhancing the academic rigor of this dissertation. Always a reliable source of expert advice, Taylor has significantly influenced the experimental approaches throughout my work, offering precise and valuable guidance that has been indispensable.

Finally, my thanks to the entire Political Science department at Washington University in Saint Louis. The department's continual support and commitment to excellence in research practices and presentation have not only bolstered my academic endeavors but have also set a standard towards which I will always strive.

Each of these individuals has not only contributed to my academic growth but also profoundly shaped my approach to research and inquiry. Their impacts extend beyond the pages of this dissertation and into the core of my professional ethos.

Dominique Lockett

Washington University in Saint Louis

July 2024

To the Moon

ABSTRACT OF THE DISSERTATION

Cognitive Landscapes:

Argument Evaluations, Misinformation Corrections, and Racial Attitudes in Modern Media

by

Dominique Lockett

Doctor of Philosophy in

Washington University in Saint Louis, 2024

Professor Jacob Montgomery

This dissertation explores three critical aspects of communication that significantly impact how individuals process information and form opinions in today's polarized media landscape. While the experiments in each chapter do not directly interact, they collectively illuminate the complex interplay of cognitive biases, misperceptions, and counter-normative speech in shaping public discourse. The first chapter examines argument congruency bias, demonstrating how individuals evaluate arguments based on their alignment with pre-existing beliefs while still distinguishing between strong and weak reasoning. It also investigates the potential of objectivity priming to mitigate these biases, yielding insights into the challenges of overcoming ingrained cognitive patterns. The second chapter focuses on the effectiveness of culturally-relevant corrections in combating misinformation, particularly within minority communities. Through experiments with Latino and Black participants, it reveals that while corrections generally reduce misperceptions, the expected superiority of in-group corrections is not consistently observed. These findings suggest a nuanced dynamic where perceived expertise and credibility play crucial roles in correction effectiveness. The final chapter investigates how exposure to various types of social media comments about race influences racial attitudes. Contrary to expectations, results indicate that brief exposures do not

significantly alter deep-seated racial beliefs, highlighting the resilience of these attitudes and the limitations of social media in challenging entrenched norms. Together, these studies contribute to our understanding of how individuals navigate the complex information environment of the digital age. They underscore the persistence of cognitive biases, the challenges in correcting misinformation, and the need for more nuanced investigations into how social media interactions shape individual and collective attitudes toward race. These findings have important implications for designing effective interventions to promote critical thinking, combat misinformation, and foster more inclusive public discourse in an era of rapid information dissemination and polarization.

Chapter 1

Beyond Bias: How do Argument Quality and Objectivity Interventions Impact Motivated Reasoning?

A significant body of recent work demonstrates the effects of motivated reasoning on individuals' capacity to accurately identify and resist misperceptions and accept true factual claims. Yet, focusing on facts alone is not sufficient for understanding the spread of misinformation. Even when individuals accept true claims, they are still vulnerable to arriving at false conclusions if they also accept flawed logic. In this paper, I focus not on how motivated reasoning leads individuals to accept false assertions, but on whether it can also lead them to accept weak logical arguments. In two survey experiments (n=1006 and n=1003), I show that individuals can distinguish between strong and weak arguments, but at the same time they are biased toward logical statements with conclusions that are consistent with their pre-existing preferences. I show that this argument congruency bias holds for both strong

and weak arguments, political and non-political topics, and multiple issue areas. In Study I, the objectivity treatment showed some effectiveness in reducing ACB, especially for weaker arguments. Unlike the first study, the second did not replicate the success in reducing ACB, indicating that the method of priming for objectivity might be critical for its effectiveness. This chapter contributes to the broader literature on motivated reasoning and misinformation by highlighting how biases extend beyond factual inaccuracies to include logical reasoning processes. By demonstrating that motivated reasoning can influence evaluations of argument strength, this research provides insights into the challenges of promoting critical thinking and rational debate in a highly polarized informational environment.

1.1 Introduction

Since 2016, a cascade of research on misinformation has shown how motivated reasoning can lead individuals to accept false or inaccurate factual claims (Corr et al. 2019; Hameleers and van der Meer 2019). Further, these misperceptions can affect political attitudes and even behaviors that have substantial impacts on the modern political landscape(James and Van Ryzin 2016; Guess and Lerner 2020).

Accepting facts, however, does not always lead to correct conclusions, particularly when flawed logic is involved. Logic is the method we use to draw conclusions from certain facts or premises. In political science, for instance, we might observe that democratic countries rarely go to war with each other, a phenomenon known as the democratic peace theory (Russett 1993). We might then conclude that promoting democracy globally would reduce international conflicts.

However, errors can occur if logic is applied incorrectly. For example, someone might argue that since democratic countries rarely fight each other, then any country not currently at

war must be a democracy. This is an example of flawed reasoning, specifically the fallacy of affirming the consequent. The observer incorrectly assumes that the absence of war is sufficient to classify a country as democratic, ignoring other factors that influence both regime type and conflict. This example demonstrates how even correct facts (democratic countries rarely go to war with each other) can lead to wrong conclusions (any peaceful country must be a democracy) if the reasoning process is flawed.

While much of the existing literature has focused on the acceptance of false factual claims, this paper shifts attention to the evaluation of argument quality, particularly how individuals assess strong versus weak arguments. A strong argument in this context is defined as one whose conclusion logically follows from its premises, whereas a weak argument is characterized by logical fallacies that undermine its validity. At the core of this investigation is the concept of argument congruency bias (ACB), which I define as the tendency to favor arguments that align with one's pre-existing beliefs, regardless of their logical strength.[1]

This paper builds on the research of scholars like Lodge and Taber (2006) and Groenendyk and Krupnikov (2020), who have explored how pre-existing beliefs affect perceptions of argument strength. Extending these findings, this study examines the impact of argument strength on motivated reasoning and investigates potential interventions to mitigate such biases. Specifically, I address three key questions. First, can individuals differentiate between strong (logically consistent) and weak (logically flawed) arguments?? Second, if individuals can make this distinction, are their evaluations influenced by motivated reasoning, leading them to prefer arguments that support their pre-existing beliefs? Third, if individuals do demonstrate ACB, is it possible to counteract this tendency by encouraging a mindset of

[1]Throughout this paper, flawed logic or logical fallacies refer to *informal fallacies*, which are errors in reasoning or content, not logical form. Examples include ad hominem (attacking the person), straw man (misrepresenting an argument), and appeal to authority (asserting a claim is true because an authority believes it).

objectivity? These questions aim to clarify the dynamics of argument evaluation and the potential for overcoming biased reasoning.

The study of motivated reasoning and its impact on individuals' ability to discern strong from weak arguments raises important questions about cognitive biases in processing information. Initially, it's crucial to establish whether individuals can reliably identify arguments that are designed *a priori* to be strong (logically consistent) or weak (logically flawed). This consideration leads to further inquiry into whether motivated reasoning can fully extinguish such distinctions in respondents' minds, and if these biases persist across various topics.

Previous research has extensively explored how factual misperceptions can be mitigated. Studies have demonstrated that priming accuracy goals can partially overcome these misperceptions (Bolsen et al. 2014; James and Van Ryzin 2016; Lodge and Taber 2013; Prior et al. 2015; Nir 2011; Slothuus and De Vreese 2010). Building on this, Groenendyk and Krupnikov (2020) found that priming goals of "open-mindedness" can also reduce biased perceptions of arguments. Despite these findings, an unexplored area remains: the potential effects of priming goals of *objectivity*. This approach is a close analogy to accuracy goals used in studies of misperceptions, suggesting a promising avenue for research into whether such priming can achieve similar reductions in biased argument perception (Custers and Aarts 2010; Marien et al. 2012; Smeesters et al. 2010).

In this paper, I present the results from two survey experiments (n=1006 and n=1003) designed to answer these questions. I extend previous work by Taber et al. (2009) and Groenendyk and Krupnikov (2020) by asking respondents to rate **both** "strong" arguments, whose conclusions follow from their premises, and "weak" arguments, those based upon logical fallacies. I show that individuals can indeed distinguish between intentionally designed strong

and weak arguments, but at the same time, they are biased toward logical statements yielding conclusions consistent with their pre-existing preferences.

I show that this *argument congruency bias* holds for both strong and weak arguments. I further explore whether such biases can be reduced by priming objectivity goals, in the same way that factual biases can be reduced by activating accuracy goals. The interventions yielded mixed results with the first indicating that priming individuals to be objective may result in a more accurate rating of weak arguments whereas the second yielded null results.

My project contributes three broad insights relevant to the scholarly discussion of motivated reasoning and political discourse. First, my results show that bias in the evaluation of arguments that favor respondent's prior beliefs is widespread, affecting evaluations of both strong and weak arguments as well as political and non-political topics. Second, I demonstrate that individuals' biases influence – but do not overwhelm – their ability to accurately rate the quality of an argument. That is, the public is not willing to value weak fallacious arguments over strong evidence-based arguments even in pursuit of directional goals.

Finally, I explore the effectiveness of a novel intervention designed to improve argument quality discernment by priming the competing goal of objectivity. My first study suggests that encouraging individuals to engage objectivity as an important principle has the potential to reduce their tendency to rely on their ideological dispositions as criterion for rating arguments. In all, my research contributes to an important stream of recent research into how individuals rate arguments of varying quality and the efficacy of priming alternative goals as a tool for reducing motivated reasoning in this context.

1.2 Motivated Reasoning

At its core, motivated reasoning is the psychological mechanism that guides individuals to process information in a way that favors their pre-existing beliefs and values (Kunda 1990). This often unconscious bias shapes how information is received, interpreted, and integrated into one's belief system. It's not merely a case of individuals preferring agreeable information but rather an active process where the brain seeks out and places more weight on evidence that confirms its hypotheses, while simultaneously discounting or ignoring evidence to the contrary (Bolsen et al. 2014; Lodge and Taber 2013; Prior et al. 2015).

Consider the example of climate change: an individual with strong environmental values might uncritically accept a report on the dire consequences of global warming, whereas a person with significant investments in the fossil fuel industry might dismiss the same report as flawed or exaggerated (Flynn et al. 2017). Here, motivated reasoning can lead to two well-informed individuals arriving at entirely different interpretations of the same data, each skewed by their respective motivations (Taber et al. 2009; Lodge and Taber 2006).

Motivated reasoning can be further broken down into two principal drivers: the pursuit of accuracy and the pursuit of specific directional goals (Lodge and Taber 2006). Accuracy-driven reasoning is marked by a genuine intent to arrive at truthful conclusions, while directionally motivated reasoning is characterized by an aim to validate pre-held beliefs. For instance, a person may seek out a broad range of news sources to understand an issue better (accuracy-driven) or selectively tune into a particular news channel that echoes their political ideology (directionally driven) (Kunda 1990; Redlawsk et al. 2010).

1.2.1 Argument Congruency Bias

The implications of motivated reasoning extend to the evaluation of arguments, giving rise to what I refer to as argument congruency bias (ACB). ACB describes a scenario where individuals assess arguments not on their logical merits but based on the conclusion's alignment with their existing beliefs (Groenendyk and Krupnikov 2020). An argument about universal healthcare, no matter how well-founded, is likely to be less persuasive to an individual opposed to government intervention in the market, whereas the same argument would resonate strongly with someone who believes in universal healthcare (Taber et al. 2009; Lodge and Taber 2013).

This bias can be influenced by two significant factors: the perceived strength of the argument and its topic (Groenendyk and Krupnikov 2020). An argument's strength is often judged subjectively; what appears cogent to one might be seen as weak to another based on their beliefs. For example, an argument that utilizes a wealth of statistical evidence to support gun control measures might be deemed strong by an advocate for stricter gun laws, while a gun rights supporter might find the same argument weak due to the interpretation of those statistics or skepticism about the source (Lodge and Taber 2006; Bisgaard 2015).

The topic of an argument also plays a crucial role in ACB. Political issues, by their nature, evoke stronger emotional responses and therefore are more susceptible to ACB. When Groenendyk and colleagues examined political arguments, they found that emotional attachment to the topic made participants more likely to judge arguments in favor of their pre-existing beliefs as stronger, regardless of the argument's actual logical consistency (Groenendyk and Krupnikov 2020).

Moreover, ACB is not limited to the realm of politics. It can influence evaluations across a wide range of issues. Take, for example, the medical field, where a patient's belief in alternative

medicine could lead them to favor arguments against vaccinations, despite overwhelming scientific evidence supporting their efficacy and safety (Nyhan and Reifler 2015; Pennycook and Rand 2019). Similarly, in the judicial system, a jury member's personal beliefs about a defendant can influence their interpretation of the evidence and arguments presented in court, potentially leading to biased judgment (Lodge and Taber 2016; Flynn et al. 2017).

Motivated reasoning and argument congruency bias represent significant challenges in critical thinking and decision-making processes. These biases can lead to the entrenchment of beliefs, polarization of opinions, and the dismissal of credible information that could otherwise inform more nuanced and balanced perspectives. Recognizing the impact of these biases, scholars have explored interventions aimed at promoting accuracy in reasoning, providing a foundation for the development of strategies to mitigate their effects.

1.2.2 Accuracy Interventions

Research regarding the broader concept of motivated reasoning has explored numerous methods to decrease biases for outcomes such as factual beliefs. Successful interventions include direct requests for accuracy and monetary incentives (Hill 2017; Prior et al. 2015). Prior et al. (2015), for instance, asked respondents about the state of the economy under the Bush and Obama administrations. They find biases in reporting based on one's partisan affiliation, but those participants who offered money or received an explicit request for accurate responses answered correctly at a higher rate. Hill (2017) had similar success when he attempted to understand how individuals integrated new information into their beliefs. In a quiz-like experiment, Hill (2017) found that prior beliefs led to bias, but a small monetary incentive could eliminate biased reporting of political facts.

Groenendyk and Krupnikov (2020) explore interventions capable of reducing motivated reasoning in the evaluation of arguments. They sought to reduce argument congruency bias by presenting half of their participants with a fictitious study claiming that open-mindedness is associated with success. The authors find slightly less bias in the average ratings of arguments among those who were in the open-mindedness treatment condition (Groenendyk and Krupnikov 2020).

Building upon this foundation of research on motivated reasoning and interventions, this study aims to extend our understanding of argument evaluation in the context of political discourse. While previous work has illuminated the impact of motivated reasoning on factual beliefs and the potential for accuracy-driven interventions, there remains a gap in our knowledge regarding how individuals assess argument quality, particularly when confronted with both strong and weak arguments. Furthermore, the potential of objectivity priming as an intervention strategy in this context remains largely unexplored. By addressing these areas, this research seeks to provide a more comprehensive picture of how motivated reasoning influences argument evaluation and to explore novel approaches to mitigating these biases. This investigation not only builds on existing scholarship but also offers new insights into the complex interplay between cognitive biases, argument strength, and the effectiveness of objectivity-focused interventions in political communication.

1.3 Theory

Existing literature on motivated reasoning has predominantly focused on the impact of pre-existing beliefs on the acceptance of factual information, demonstrating that biases in information processing can be mitigated through accuracy priming (Bolsen et al. 2014; James and Van Ryzin 2016; Lodge and Taber 2013; Prior et al. 2015; Nir 2011; Slothuus and

De Vreese 2010). Extending this research, Groenendyk and Krupnikov (2020) explored how open-mindedness could similarly reduce biased perceptions. However, the novel aspect of this study lies in its construction of *weak arguments* for participants to asses and a focus on whether priming for *objectivity* could achieve comparable results in the realm of argument evaluation, a hypothesis less explored in existing research (Custers and Aarts 2010; Marien et al. 2012; Smeesters et al. 2010).

This study expands the literature on motivated reasoning by analyzing how people respond differently to strong and weak arguments. Using insights from psychology and political science, I seek to not only confirm the presence of bias but also explore the conditions that might enhance or mitigate it. I specifically address how individuals differentiate between arguments that are logically coherent and those that are weak because they're based on logical fallacies. The implications of argument strength on ACB are significant, suggesting that stronger arguments might reduce bias by prompting more thorough consideration, whereas weaker ones could exacerbate it by encouraging rationalizations that align with existing beliefs.

1.3.1 Research Questions

The central question posed is whether individuals inherently favor arguments that align with their existing beliefs, thus demonstrating a preference that might skew rational evaluation. This question is pivotal for understanding how deeply ingrained biases can influence not just the reception of factual information but also the critical evaluation of argumentative strength and logical coherence in discourse.

Research question 1: Is argument congruency bias observable in the evaluations of arguments?

Research Question 2 extends this exploration by examining whether individuals can discern the strength of arguments, assessing whether they rate stronger arguments higher than weaker ones. In this context, "weak" arguments are those that are based on informal logical fallacies, meaning their content does not sufficiently support their conclusions.

For example, a weak argument I formulated for this study was, "*If we allow the government to regulate the use of guns, soon they will be forcibly removing guns from our homes*." This argument is an example of the "slippery slope fallacy", which suggests that a small step in one direction will inevitably lead to extreme consequences.

Compare this to the corresponding strong argument constructed by Taber and Lodge (2006a): "*A main reason why our murder rate is so high is that most crime victims do not resist. These victims are twice as likely to be injured compared to those who defend themselves. Carrying a gun is thus one's ultimate protection against violent crime*."

The strength of the second argument lies in its use of specific claims and a logical chain of reasoning. It presents a premise (non-resistance leads to higher injury rates), provides supporting information (comparative injury statistics), and draws a conclusion based on this information. This structure makes the argument more convincing and rational. The first argument, on the other hand, is speculative and relies on an unsupported assumption that one action will inevitably lead to extreme consequences, without providing any intermediate steps or evidence to support this claim.

Research question 2: How accurately do individuals distinguish between strong and weak arguments?

This progression is vital for understanding if the natural bias towards congruent arguments can be counterbalanced by the intrinsic quality and logical strength of the arguments presented. This inquiry not only deepens our understanding of bias in argument evaluation but also tests the potential of critical thinking to mitigate these biases.[2]

1.3.2 Hypotheses

Hypothesis 1 focuses on the presence and impact of argument congruency bias, testing whether interventions that promote objectivity can effectively reduce this bias. The hypothesis is evaluated using a between-subjects design, inspired by prior research that demonstrates how accuracy interventions, like direct accuracy requests or monetary incentives, can decrease biases in how information is processed. This approach is grounded in the idea that activating thoughts on objectivity in evaluators can lead to more balanced and accurate assessments of arguments, thereby countering the influence of their pre-existing beliefs.

Building on this, Hypothesis 2 examines whether stronger arguments, which demand more critical engagement, can further decrease the biases identified in Hypothesis 1. This inquiry explores how the robustness of an argument influences cognitive processing and bias mitigation, aiming to clarify how deeper analytical engagement affects evaluative biases.

Hypothesis 1: I think about objectivity, therefore, I am objective

Hypothesis 1 examines how fostering a mindset of objectivity may counteract ACB– the tendency to favor information that aligns with one's pre-existing beliefs.

[2]See page 120 of Appendix A for details on a third research question exploring the impact of topic on participants' evaluations.

Hypothesis 1: Individuals exposed to the objectivity treatment will have less argument congruency bias than those who are not.

The mechanism underlying this treatment is automatic goal pursuit (Bargh et al. 2001; Custers and Aarts 2010; Hassin et al. 2009; Marien et al. 2012; Strack and Deutsch 2011). The theory of automatic goal pursuit suggests that the activation of an idea can lead to the unconscious pursuit of said idea as a goal. This theory suggests that "no conscious intervention, act of will, or guidance is needed ... [to cause] goal pursuit" (Bargh et al. 2001; Custers and Aarts 2010; Hassin et al. 2009; Marien et al. 2012). Research in the field of psychology has even shown that goals activated unconsciously operated as effectively as when individuals consciously decide to pursue goals (Marien et al. 2012). One example of this effect involves a method commonly used in social cognition research.

In the study of automatic goal pursuit, a compelling method involves the use of sentence unscrambling tasks to prime specific cognitive and behavioral goals. Participants are asked to unscramble a sentence and psychologists choose specific words which are used to prime participants toward various behaviors. Gollwitzer and Bargh (2005) performed such a test and provided half their sample with words about memorization and the other half with words about evaluation. Respondents then performed a second task in which they read about a hypothetical person and were later asked to report on aspects of what they had read (Gollwitzer and Bargh 2005). The researchers find that those who unscrambled sentences with words about memorization did a better job recalling information from the second task. This suggests that the participants in the memorization treatment group had automatically set goals about memorization and pursued those goals in the second task (Gollwitzer and Bargh 2005).

The principle of automatic goal pursuit suggests that it is possible to design interventions that interrupt the automatic biases in people's thought processes. Motivated reasoning and

its subset argument congruency bias are phenomena that occur because of automaticity. Automaticity is a 'knee-jerk' emotional response. When exposed to stimuli-related concepts such as feelings, then dispositions, facts or beliefs are instantaneously activated (Lodge and Taber 2013). Because our *reasoning* is *motivated* by our emotions, we seek to justify our emotions by constructing logical grounds to support them. Automaticity suggests that our emotions determine the type of deliberation that occurs ("I disagree with this conclusion; I will think about all the ways it can be disproved"). Moreover, prior work suggests that this process can be overridden if an individual is motivated to be accurate (Groenendyk and Krupnikov 2020; Hill 2017; Prior et al. 2015; Redlawsk et al. 2010). I aim to interrupt the the automatic pursuit of directional goals by activating the competing goal of objectivity. Drawing on the theory of automatic goal pursuit, I argue that such activation can guide individuals toward more objective behavior. That is, rather than constructing rationalizations for their feelings, individuals will set aside their feelings to provide less biased evaluations.

Hypothesis 2: To think is to be biased

Following the exploration of interventions to mitigate argument congruency bias, Hypothesis 2 shifts the focus to the role of argument strength in influencing evaluative biases. Research suggests that individuals engage in more extensive cognitive processing when confronted with strong arguments, particularly those that conflict with their pre-existing beliefs (Groenendyk and Krupnikov 2020; Hart and Nisbet 2012; Lodge and Taber 2013, 2006; Flynn et al. 2017). This increased engagement, however, may not always lead to reduced bias. Instead, it might provide more opportunities for motivated reasoning, allowing individuals to scrutinize and potentially rationalize their pre-existing views more thoroughly (Kunda 1990; Redlawsk et al. 2010).

Hypothesis 2: Stronger arguments will elicit more argument congruency bias than weaker arguments, as they provide more opportunities for motivated reasoning and belief reinforcement.[3]

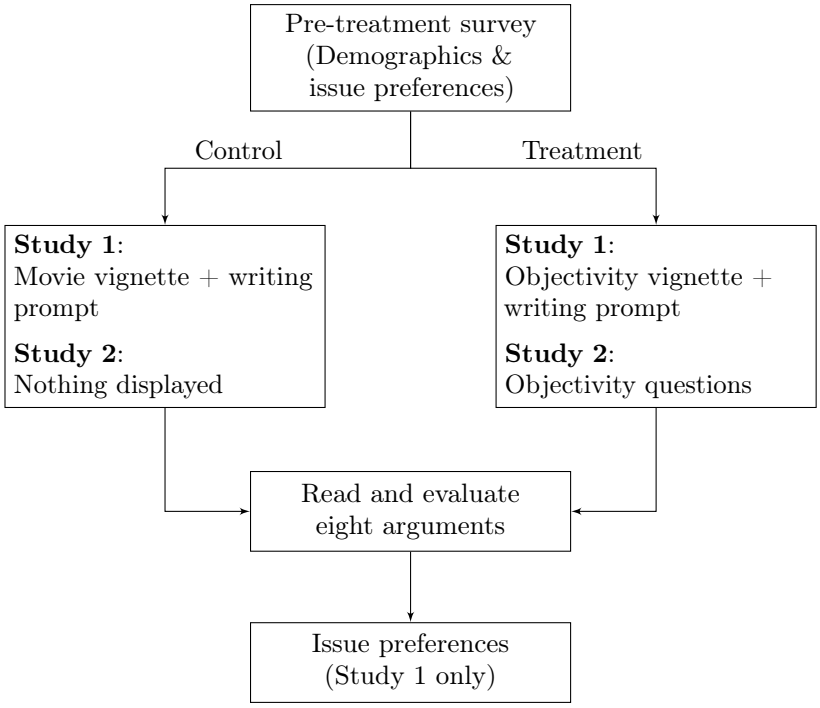
Strong arguments, with their logical coherence and plausibility, may paradoxically reinforce ACB by offering more points of engagement for individuals to align with their existing beliefs (Taber et al. 2009; Lodge and Taber 2006). The depth and complexity of strong arguments could provide more material for individuals to selectively interpret or counter-argue, potentially strengthening their original positions (Bisgaard 2015; Nyhan and Reifler 2015).

Conversely, weak arguments, often based on logical fallacies or insufficient justifications, might elicit less cognitive engagement (Pennycook and Rand 2019). When these weak arguments align with an individual's existing beliefs, they may be accepted with less scrutiny. However, when they conflict with pre-existing views, their obvious flaws might make them easier to dismiss, potentially resulting in less overall bias (Lodge and Taber 2016; Flynn et al. 2017).

The interplay between argument strength and ACB underscores the complex relationship between cognitive biases and evaluative standards (Bolsen et al. 2014; Prior et al. 2015; Nir 2011; Slothuus and De Vreese 2010). While critical thinking and objectivity remain crucial for balanced assessments, the findings suggest that merely increasing cognitive engagement through stronger arguments may not be sufficient to reduce bias (Custers and Aarts 2010; Marien et al. 2012; Smeesters et al. 2010). These dynamics highlight the need for nuanced approaches in promoting critical thinking and objectivity, particularly in contexts where polarized opinions and confirmation bias are prevalent (James and Van Ryzin 2016; Lodge and Taber 2013; Taber et al. 2009).

[3]This hypothesis represents a revision from the pre-registered hypothesis, which stated: "The treatment will be more effective when individuals are evaluating arguments that are weak." See page 133 of Appendix A.

Figure 1.1: Flowchart of experimental design



Participants complete a pre-treatment survey, then are randomly assigned to control or treatment conditions. In Study 1, the control group writes about a movie, while the treatment group writes about the value of objectivity. In Study 2, the control group receives no display, while the treatment group answers questions about objectivity. All participants then evaluate eight arguments. For full details of the treatments and survey instruments, see page 130 and page 140 of Appendix A

To explore my research questions and hypotheses, I constructed and distributed two survey experiments in June 2019 and May 2020. The first relied on a convenience sample provided by Amazon's Mechanical Turk (MTurk), and the second was a nationally representative survey hosted on NORC at the University of Chicago. You can refer to Figure 1.1 for an understanding of the survey flow of the two studies.

In Study 1, participants were asked to engage in a reflective exercise by writing a short paragraph on the importance of objectivity in both public and personal contexts. This task was designed to activate their internal goal of being objective.

In contrast, Study 2 employed a less intensive approach, where participants were presented with five statements about the significance of objectivity and asked to rate their agreement with each statement. This method aimed to prompt participants to consider the value of objectivity without requiring them to articulate their thoughts in writing.

Both studies utilized a randomized control trial (RCT) design to ensure that any observed effects could be attributed to the intervention rather than other confounding factors. Participants were randomly assigned to either the control group, which received no specific treatment related to objectivity, or the treatment group, which was exposed to the objectivity-promoting intervention. Following the intervention, all participants were asked to read and evaluate a series of arguments on various topics. The key measure of interest was whether the treatment group exhibited less argument congruency bias compared to the control group.

The remainder of this paper will expand upon the details and differences of the two experimental designs. In separate sections, I will discuss each studys' research design and results before turning to a unified discussion on the findings and limitations of both and, finally, a conclusion.

1.4 Study I

The first study investigated the presence of argument congruency bias and its potential mitigation through the promotion of objectivity. In this study, I explored how individuals evaluate arguments of varying strengths, particularly when these arguments align or conflict with their pre-existing beliefs. By examining the impact of the objectivity intervention, I sought to understand whether priming objectivity goals would reduce biases in the evaluation of arguments. The findings from this study will contribute to the overall understanding of

motivated reasoning and the effectiveness of interventions designed to encourage objective assessment of arguments.

1.4.1 Research Design

To achieve these goals, an online experiment was conducted using Amazon's Mechanical Turk (MTurk) platform in June 2019. This platform, widely used and validated in social science research (Berinsky et al. 2012; Mullinix et al. 2015; Coppock 2019), enabled the recruitment of 1112 respondents who were randomly assigned to either a treatment or a control group. The participants were tasked with evaluating four arguments on the topic of gun control, categorized by their strength as either strong or weak arguments.[4] Refer to Figure 1.1 for a visual understanding of the survey flow.

The study specifically aimed to examine how the strength of an argument (strong vs. weak) influences participants' evaluations, especially when these arguments align or conflict with their pre-existing beliefs (congeniality). To measure congeniality and study argument congruency bias, participants first completed a pre-treatment survey where they provided demographic information and their stance on gun control, rated on a scale from 1-5, with higher values indicating more support for gun control measures. Based on these ratings, participants were categorized as either pro- or anti-gun control.

They were then presented with four arguments about gun control, classified as either strong or weak.[5] The study assessed how participants rated these arguments, particularly observing if they rated arguments that aligned with their pre-existing beliefs (congenial arguments) higher than those that did not (uncongenial arguments).

[4]Gun control was selected as the focus due to its significant relevance in American politics and its established use in prior research by Taber and Lodge (2006a).

[5]See page 131 of Appendix A for exact arguments that were evaluated.

This approach allowed me to evaluate the presence and extent of argument congruency bias, examining whether participants' evaluations were influenced more by the strength of the arguments or by their own pre-existing beliefs.

Treatment groups

The experimental design consisted of two primary groups: a control group and an objectivity treatment group. Both groups were exposed to a reading and writing exercise.

Table 1.1 provides the reading prompts for both groups. The treatment group was introduced to the concept of objectivity, while the control group read a paragraph about movies, chosen for its neutrality and lack of direct relevance to the contentious subjects of gun policy.

Table 1.1: Reading prompts for control and treatment groups

Control Group Prompt	**Treatment Group Prompt**
Movies can be fun, but don't underestimate how much they can provide to our society. Movies encourage ideas and social commentary within communities. They have the power to express a culture's ideals and shape them. Movies are important because they give us the ability to form lasting human connections by letting us share our experiences with each other.	Objectivity is the ability to make judgments without relying on personal feelings or opinions. Being objective ceans applying the rules fairly and treating everyone the same rather than showing favoritism. Objectivity requires you to consider perspectives other than your own to achieve less biased conclusions.

Presents the initial reading prompts: the control group discusses the societal impact of movies, while the treatment group focuses on the importance of objectivity in judgment. Subsequently, the control group was asked to "Please write a brief paragraph explaining why objectivity is valuable to you and society." While the treatment group was asked to "Please write a brief paragraph explaining why objectivity is valuable to you and society." For further details on the survey instrument, see page 130 of Appendix A.

Following the reading phase, participants in the treatment group were asked to write a brief paragraph explaining why objectivity was important in their lives and for society. This

exercise aimed to bring objectivity to the forefront of their thoughts and reinforce the notion that objectivity is a desirable trait. Table 1.2 showcases three exemplary responses from participants in the objectivity treatment group. Their written responses were then displayed at the top of each page during the subsequent argument evaluation task.

In contrast, participants in the control group were prompted to write about the last movie they watched. The writing prompts for both groups had a minimum word count requirement to ensure a comparable level of engagement. The following sections will expand on the data analysis procedures, and the results obtained.

Table 1.2: Responses to the treatment writing prompt

No.	Participant Responses
1	"It is important to have everybody's voice and opinion heard. Society is comprised of many different individuals of different ethnicities and backgrounds. Thus, the people of the government should use objectivity to determine what is right for the people as a whole as opposed to a small group of people."
2	"Being able to take in all information and points of view allows for compromise and progress to be made. It is the only way to build coalitions and please as many people as possible even though it might be impossible to give exactly what some wanted."
3	"Objectivity is important to both me and to society because it's important to think about many sides of an issue, and not just your own. It's OK to disagree with someone or to not agree with a way of thinking, but it's important to have informed opinions about issues that are important to society as a whole."

Displays selected responses from participants asked to write about the importance of objectivity in society. These excerpts illustrate various perspectives on how objectivity influences societal decision-making and personal thought processes. See page 119 for a random sample of responses from the objectivity treatment group.

Variables

The primary dependent variable in this study is argument congruency bias, calculated for each of the four argument categories as the difference between evaluations of congenial and

uncongenial arguments. To construct the ACB measure, I utilized participants' pre-treatment assessments on gun control and pineapple on pizza. Their responses to these questions allowed us to categorize their positions as either pro or anti for each issue.

Participants were then assigned to the respective stance categories for each argument. For example, if a participant indicated a pro-gun control stance, they would be categorized as having a congenial stance for pro-gun control arguments and an uncongenial stance for anti-gun control arguments.

Calculating argument congruency bias

To calculate the ACB for each participant, we subtracted the scores given to congenial arguments from the scores given to uncongenial arguments. A positive ACB score indicates a bias toward congenial arguments, while a negative score suggests a bias toward uncongenial arguments. We aggregated these scores across all categories to derive an overall ACB measure for each participant.

To quantify the effectiveness of these interventions, the primary focus is on measuring the average treatment effect (ATE). This is achieved through linear regression analysis with clustered standard errors, which allows for a robust estimation of the impact of the treatment on reducing argument congruency bias. The average treatment affect for Study 1 (ATE_1) is calculated using the following formula:

$$ATE_1 = \underbrace{\underbrace{\left[\bar{y}_{(c,T)} - \bar{y}_{(u,T)}\right]}_{\text{Bias in treatment}} - \underbrace{\left[\bar{y}_{(c,R)} - \bar{y}_{(u,R)}\right]}_{\text{Bias in control}}}_{\text{Average treatment effect}} \tag{1.1}$$

In this equation, c denotes congenial arguments (those aligning with participants' prior beliefs), and u denotes uncongenial arguments. T represents individuals in the treatment condition, while R signifies those in the control condition.

This calculation provides an insight into the overall effectiveness of the treatment in mitigating bias. Taking the average across all arguments is essential to ensure that the results are not skewed by any single argument and to provide a comprehensive measure of bias reduction across different argument qualities and conditions.

To further refine the analysis, a more complex difference-in-difference-in-differences design is employed to account for the variation in argument quality presented to participants. This approach is captured in the following equation:

$$ATE_2 = \underbrace{\underbrace{\left[\bar{y}_{(c,T,H)} - \bar{y}_{(u,T,H)}\right]}_{\substack{\text{Bias in treatment with} \\ \text{high quality argument}}} - \underbrace{\left[\bar{y}_{(c,R,H)} - \bar{y}_{(u,R,H)}\right]}_{\substack{\text{Bias in control with} \\ \text{high quality argument}}}}_{\text{Average treatment effect among high quality arguments}} - \underbrace{\underbrace{\left[\bar{y}_{(c,T,L)} - \bar{y}_{(u,T,L)}\right]}_{\substack{\text{Bias in treatment with} \\ \text{low quality argument}}} - \underbrace{\left[\bar{y}_{(c,R,L)} - \bar{y}_{(u,R,L)}\right]}_{\substack{\text{Bias in control with} \\ \text{low quality argument}}}}_{\text{Average treatment effect among low quality arguments}} \tag{1.2}$$

In Equation (1.2), H indicates arguments of high quality (strong arguments) and L indicates arguments of low quality (weak arguments). This equation will assist in better understanding Hypothesis 2 and the ways in which the strength of the argument moderates one's bias.

For the triple difference-in-differences design to be valid, it assumes that in the absence of treatment, the difference in bias between strong and weak arguments would follow the same trend in both the treatment and control groups (parallel trends assumption). This ensures that any observed changes can be attributed to the intervention rather than other confounding factors. Additionally, it is assumed that participants did not change their behavior before the

treatment due to expecting the intervention (no anticipation effect), and that the intervention impacts all individuals within the treatment group consistently (homogeneous treatment effect). These assumptions help ensure that any observed changes can be attributed to the intervention rather than other confounding factors and are assumed to be upheld in the sterile survey environment.

Additional variables

The main independent variable is the treatment condition (objectivity or movies writing prompt), along with the strength of the argument being evaluated (strong or weak). The data was transformed from 1006 (1112 minus those who did not pass the attention check) participants to 2012 observations, allowing for individual assessments for both congenial and uncongenial opinions. To account for the change in dimensions, clustered standard errors are used to address the issue of non-independence within the data. Since each participant provided multiple evaluations (one for a congenial argument and one for an uncongenial argument), these evaluations are not independent of each other. Clustering standard errors by participant helps to adjust for this within-subject correlation, ensuring that the standard errors accurately reflect the variability in the data.

Control variables include participants' political party preference, strength of partisanship, ideology, education level, income, and race. Additionally, the study incorporates variables measuring participants' scores on the need for cognition and need to evaluate scale, which assess the extent to which individuals engage with and evaluate information, a factor relevant to argument evaluation. Individuals with a higher propensity for thoughtfulness are often more likely to employ motivated reasoning and exhibit argument congruency bias, as they have a greater ability to generate counter-arguments and evidence (Bizer et al. 2000; Nir 2011).

The survey included an attention check within a matrix table presenting statements related to these traits. Over 90% of participants in each group passed this attention check, and those who failed were excluded from the analysis. Results including participants who failed the attention check can be found in Appendix A, Table A.5. The survey experienced minimal attrition, with only six of the 1012 participants failing to complete it.

1.4.2 Results

The randomized assignment of participants into the two groups achieved equivalent averages across demographic variables such as age, income, ideology, as well as the average opinion about gun control and pineapples on pizza. The exception to this balance is the gender of participants in the treatment and control. While the treatment included 50% males, the control only had 44% males. Overall, about 83% of respondents supported gun control.[6]

Research questions

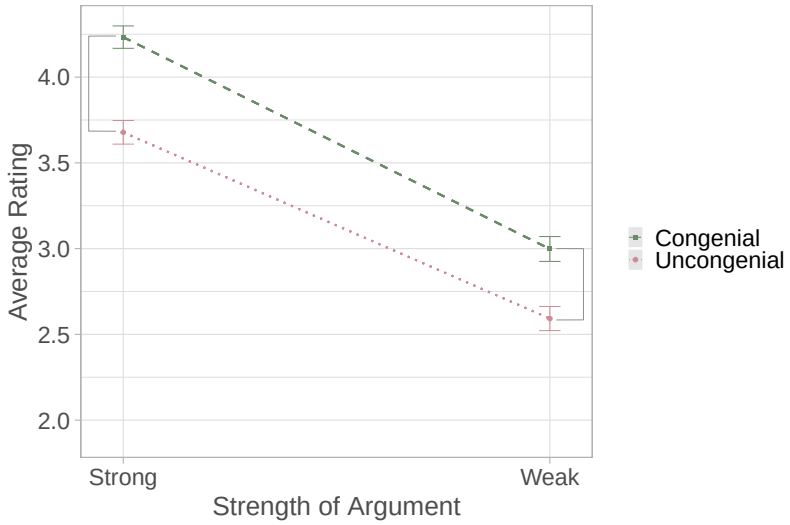
The concept of argument congruency bias suggests that individuals evaluate arguments based on how well they align with their pre-existing beliefs. The first research question seeks to validate the presence of argument congruency bias, building on the findings of Lodge and Taber (2013) and Groenendyk and Krupnikov (2020). The second research question explores whether individuals can distinguish between strong and weak arguments.

Figure 1.2 illustrates the difference in ratings between congenial and uncongenial arguments, taking into account the strength of the arguments. This comparison helps us understand the extent of argument congruency bias and whether individuals can differentiate between strong and weak arguments. If argument congruency bias exists, we expect a significant difference in

[6]See page 117 of Appendix A for exact values.

ratings between congenial and uncongenial arguments, regardless of their strength. Conversely, if there is no bias, the ratings for congenial and uncongenial arguments should be similar.

Figure 1.2: Evidence of argument congruency bias and argument strength differentiation



Shows the mean ratings of arguments classified by their strength (strong vs. weak) and their congeniality (congenial vs. uncongenial). Point estimates include 95% confidence intervals. Grey brackets indicate the level of argument congruency bias, highlighting the difference in ratings between congenial and uncongenial arguments. The figure provides evidence that individuals rate congenial arguments higher than uncongenial ones and can distinguish between strong and weak arguments. See Figure 1.4 for Study II outcomes.

The results in Figure 1.2 confirm that argument congruency bias is indeed observable. On average, weak arguments received a 0.32-point higher rating when they were congenial to participants' prior beliefs, while strong arguments experienced a slightly higher bias, with congenial arguments rated 0.6 points higher. Given the 1-5 Likert scale used, these differences represent a 6.4% increase in ratings for congenial weak arguments and a 12% increase for congenial strong arguments, underscoring the substantial impact of argument congruency bias on evaluations.

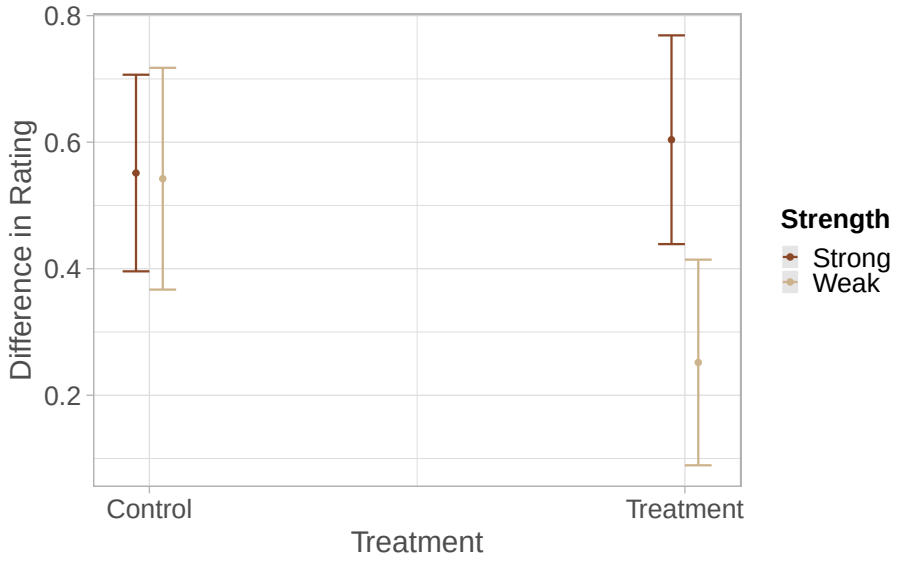
Regarding the second research question, the data indicates that individuals can indeed differentiate between strong and weak arguments. Regardless of congeniality, strong arguments received an average rating of 3.94, while weak arguments were rated around 2.73 on a 1-5 scale.

In summary, the results confirm the presence of argument congruency bias, as individuals consistently rate arguments that support their prior beliefs higher than those that do not. Additionally, the findings suggest that individuals can distinguish between strong and weak arguments, with strong arguments receiving higher ratings regardless of congeniality. These findings provide a foundation for further investigation into the ways in which biases can be mitigated.

Hypotheses

My main hypothesis posits that individuals exposed to the objectivity treatment will have less argument congruency bias than those who are not. For **Hypotheses 1**, I measure argument congruency bias as the difference between the rating of congenial arguments and uncongenial arguments. A value above 0 will indicate that individuals rate congruent arguments more highly than incongruent arguments, demonstrating that they evaluate the messages according to the *conclusion* of the argument. If there is no difference between an individuals' rating of congenial and uncongenial arguments, it suggests that individuals rate arguments based on their *strength* rather than by their congeniality to their prior beliefs. I rely on OLS to analyze the outcome of my experiment. Given the structure of the dataset, standard errors were clustered at the individual level to account for multiple occurrences of the same respondent.

Figure 1.3: Treatment decreases argument congruency bias when weak



Marginal effect of exposure to objectivity treatment on difference in rating of congenial and uncongenial arguments. Point estimates (with 95% confidence intervals) from OLS regression with clustered standard errors. Both topics are included. Control variables include gender, college education, race, ideology, party ID, age, attention to politics, need to evaluate, need for cognition, political knowledge and strength of partisanship. See Table A.5 for associated regression analyses.

Figure 1.3 illustrates the average treatment effect of my experiment.[7] The x-axis displays the treatment condition and the y-axis represents the level of argument congruency bias present in the evaluation. The results indicate that argument congruency bias is around 0.55 for strong and weak arguments. The treatment reveals that my intervention had a different effect depending on the strength of the argument. Among strong arguments, the treatment had a slight increase when moving from control to treatment suggesting more argument congruency bias among strong arguments when exposed to the treatment. Among weak arguments, a decrease in the difference in rating of congenial and uncongenial arguments of roughly .3 is observed.

[7]Regression analyses for Figure 1.3, those who failed the attention checks and for each topic separately are available in page 124 of Appendix A.

The empirical findings from Study 1 corroborate the theoretical expectations posited under **Hypothesis 2**: Stronger arguments illicit more bias from evaluators. The data indicate that the strength of an argument significantly moderates the bias present in evaluative judgments. Specifically, strong arguments were found to exhibit greater bias, aligning with the theory that deeper cognitive engagement with stronger, more convincing arguments can reinforce pre-existing biases rather than diminish them. In a control setting, where participants were not primed for objectivity, weak arguments that were congruent with participants' beliefs received ratings significantly higher—over 10% more favorable—than their uncongenial counterparts.

Interestingly, this bias towards congenial weak arguments was reduced by approximately 5% when participants underwent an objectivity-oriented treatment. This reduction suggests that the treatment may prompt participants to scrutinize weak arguments more critically, assessing them on their merits rather than through the lens of bias. However, these results also underscore the complexity of cognitive biases: even when individuals are encouraged to think objectively, stronger arguments continue to sway them more due to their depth and the cognitive effort they command.

These findings emphasize the intricate relationship between cognitive engagement, argument strength, and bias. They highlight the challenges in fostering genuine objectivity, particularly in contexts where individuals are deeply entrenched in their viewpoints. Further research is necessary to explore how different types of cognitive interventions can more effectively reduce bias, especially when individuals engage with strong arguments that resonate with their existing beliefs. This ongoing exploration is crucial for developing strategies that not only promote critical thinking but also enhance the ability to evaluate arguments impartially, irrespective of their strength.

Taken together, these results conform to the expectations outlined in the main hypothesis and provides tentative insight into possible mechanisms. Among weak arguments, I provide evidence to suggest individuals are capable of moderating their biases. This may suggest that individuals may be especially inclined to counter-argue strong uncongenial arguments because these types of arguments have the most potential to challenge the validity of one's prior beliefs. Alternatively, abandoning support for fallacious congenial arguments may be much easier when the goal of objectivity is activated.

1.5 Study II

Study I yielded promising results, leading to a subsequent investigation in Study II. This study aimed to further explore the relationship between ACB and argument strength. Several modifications were made from the original design, including, adding immigration as a second topic and removing the written portion of the treatment condition.

1.5.1 Research Design

This experiment was conducted through The American Social Survey (TASS), sponsored by the Weidenbaum Center at Washington University in St. Louis. A nationally representative sample was obtained via AmeriSpeak using a NORC National Frame address-based sample. Participants were randomly assigned to either a control or a treatment group, with the latter being exposed to the concept of objectivity.

In this study, the strong gun control arguments were modified to reference reputable sources, enhancing their credibility. Unlike the first study, attention checks were omitted, and a subset of control variables was collected separately from the core survey.

A significant alteration in Study II was the approach to the objectivity treatment. Due to logistical limitations, participants could not be compelled to complete a writing prompt. Instead, the treatment group was presented with statements about objectivity and asked to rate their agreement. This method aimed to subtly activate objectivity goals, albeit in a less direct manner than the original treatment.

Participants' issue preferences were gauged through two questions per topic.[8] Following random assignment to the treatment or control group, they evaluated eight arguments. The analysis incorporated clustered standard errors and, uniquely for Study II, included weights to ensure representativeness. The study's design and hypotheses were pre-registered with the Evidence in Governance and Politics (EGAP) registry. Details of the registration are available for review in Appendix A page 133.

Treatment groups

In Study II, the distinction between the control and treatment groups was based on an intervention involving five statements related to objectivity. The treatment group was first provided with a definition of objectivity, emphasizing its role in achieving unbiased judgment by considering perspectives beyond personal feelings and opinions. Subsequently, participants in the treatment group were presented with four statements emphasizing the importance of objectivity in various contexts. They were asked to rate their agreement with each statement on a 1-5 scale, where 5 represented strong agreement. The definition and statements were as follows:

- Objectivity is the ability to make judgments without relying on personal feelings or personal opinions. Being objective means applying the rules fairly and treating everyone the same rather than showing favoritism. Objectivity requires you to consider perspectives other than your own to achieve less bias.

[8] See page 128 of Appendix A for the wording of the questions.

- Objectivity is important because it allows people to think carefully about opinions that differ from their own.
- Being objective makes it easier for people to get as close to the truth as possible.
- Objectivity is important to society as it ensures that rules are applied fairly.
- Objectivity is important because it allows you to examine facts without letting emotions get in the way.

Variables

In Study II, control variables included indicators for gender, education level, race, ideology, and age. The variables for party ID and partisan strength were consolidated into a single measure, asking participants to position themselves on a scale ranging from strong Democrat to strong Republican.

The dependent variables in the replication study remained consistent with those in the original experiment. Argument congruency bias is defined as *the difference in rating of congenial and uncongenial arguments for each individual.* This metric quantifies the discrepancy between a participant's evaluation of arguments that align with their pre-existing beliefs (congenial) and those that do not (uncongenial).

1.5.2 Results

Randomization was successful as the control and treatment groups achieved approximately equal averages across age, ideology, income, gender, and race, as well as the average opinions toward gun control and immigration.[9] For the most part, these descriptive statistics align quite well with those of the original experiment. Much like the original experiment, participants in this study held more liberal opinions toward gun control with 78% in favor of gun control

[9]Balance outcomes can be seen in page 133 of Appendix A

and 67% in favor of immigration. One deviation between the two studies was education: on average about 55% of the MTurker's of the first study reported having a Bachelor's versus around 35% in the second study.

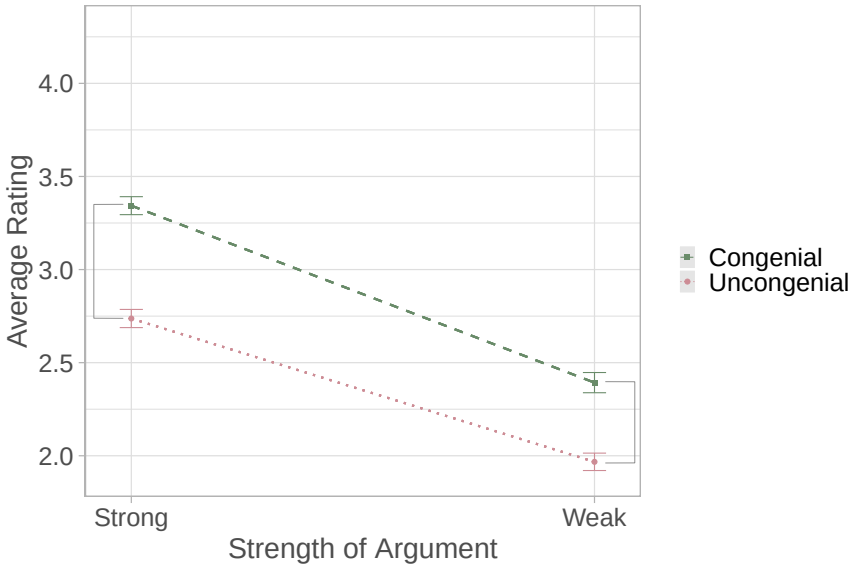
Research questions

The first research question revisits the concept of argument congruency bias, which posits that individuals prefer arguments that align with their pre-existing beliefs, rating them higher than those that do not. Figure 1.4 from Study II not only reaffirms the presence of this bias but also illustrates that it remains robust across argument strengths. The congruency effect is evident in both strong and weak arguments, with congenial arguments consistently rated higher than their uncongenial counterparts, mirroring the patterns observed in Study I. This consistency across studies highlights the pervasive nature of bias in argument evaluation, irrespective of the argument's inherent strength.

Research question 2 further explored how individuals differentiate between strong and weak arguments. While the differentiation is clear, as shown in Figure 1.4, a notable shift occurs in the overall ratings: strong arguments receive lower ratings in Study II compared to Study I, averaging closer to 3 than the 4 observed previously. This drop could be attributed to the introduction of actual data within the arguments, possibly increasing their complexity and thus the cognitive load for participants. This increased complexity might have led to more rigorous scrutiny, making participants more critical and potentially skeptical of the arguments' assertions.

Having explored how arguments are evaluated based on their strength and congeniality, the focus now shifts to experimental interventions aimed at reducing biases. This section assesses the effectiveness of objectivity treatments in a controlled experimental setup.

Figure 1.4: Evidence of argument congruency bias and argument strength differentiation

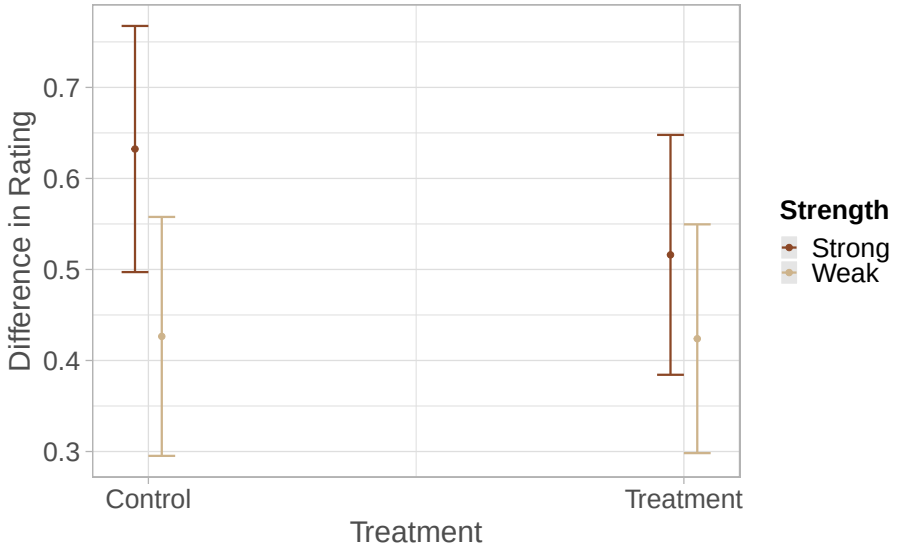


Presents a mean comparison of ratings of arguments classified by their strength (strong vs. weak) and their congeniality (congenial vs. uncongenial) in Study II. Point estimates include 95% confidence intervals. Grey brackets indicate the level of argument congruency bias, highlighting the difference in ratings between congenial and uncongenial arguments. The results show that argument congruency bias is observable and similar in magnitude to Study I. See Figure 1.2 for Study I outcomes (the same scale is used).

Hypotheses

The analysis progresses to test the hypotheses that investigate the impact of objectivity treatments on evaluative biases, aiming to uncover the underlying mechanisms that influence argument assessment.

Figure 1.5: Treatment results in an insignificant reduction of argument congruency bias among strong arguments



Shows the impact of the objectivity treatment on the difference in ratings between congenial and uncongenial arguments. Point estimates with 95% confidence intervals are derived from OLS regression with clustered standard errors and sample weights. The analysis includes control variables such as gender, college education, race, ideology, strength of partisan ID, and age. Results indicate an insignificant reduction in argument congruency bias among strong arguments and no change among weak arguments when moving from control to treatment.

Hypothesis 1 expects that individuals in the objectivity treatment will have less argument congruency bias than those in the control. The results in Figure 1.5 suggest that this may be the case. Among strong arguments, we see a modest reduction in argument congruency

bias moving from control to treatment. However, this change is not significant, which is confirmed in the regression analysis associated with Figure 1.5, which can be seen on page 138 of Appendix B. This suggests that while the treatment may influence evaluative processes, its effect is not robust enough to significantly alter the entrenched biases affecting argument evaluation.

Hypothesis 2 investigates whether the strength of the argument continues to moderate the bias in evaluating arguments in a nationally representative sample. Figure 1.5 finds no significant change in bias for weak arguments between control and treatment conditions. This outcome suggests that while argument strength is a factor in bias, the objectivity treatment's influence is limited, highlighting the challenges in mitigating bias through interventions that solely emphasize objectivity.

Study II confirms the enduring nature of argument congruency bias and elucidates the complexities involved in mitigating such biases through objectivity-focused interventions. The study's insights into the interaction between argument strength and cognitive engagement reveal that stronger, data-rich arguments do not necessarily lead to less bias but may instead invite more critical evaluation. These findings underscore the nuanced challenges in fostering objective evaluations and suggest that more comprehensive strategies may be necessary to effectively reduce biases in argument assessment.

1.5.3 Discussion

The findings from both Study I and Study II deepen our understanding of the dynamics of argument congruency bias and its interaction with argument strength. Across both studies, a clear pattern emerged: participants consistently rated arguments that align with their pre-existing beliefs higher than those that do not, regardless of the inherent strength or

quality of the arguments. This persistent preference underscores a fundamental bias favoring alignment with personal beliefs over quality. Moreover, the results indicate that stronger, data-supported arguments do not necessarily counteract this bias. Instead, they may engage participants more deeply, possibly reinforcing bias through increased cognitive scrutiny.

While Study I showed some promise in mitigating biases with an objectivity treatment that slightly reduced bias among weak arguments, suggesting a potential for deeper cognitive engagement, Study II did not replicate this effect, highlighting the challenges of altering deeply entrenched cognitive biases with subtle and brief interventions. The relative success of the objectivity treatment in Study I, compared to the lack of significant findings in Study II, raises an interesting point about participant engagement. The writing prompt in Study I likely demanded more active engagement from participants, requiring them to articulate their thoughts about objectivity. This process might have made the concept of objectivity more salient in their minds, thereby influencing their subsequent argument evaluations more effectively than the simpler agreement rating used in Study II.

In discussing the overarching limitations, the significant number of participants supporting gun control could affect the generalizability of the results, potentially biasing the findings towards those more likely to favor pro-gun control arguments. Moreover, Study I utilized an MTurk sample, which may not accurately represent the broader population, possibly limiting the applicability of the findings. Additionally, the possibility of alternative mechanisms, such as expressive responding or social desirability bias, cannot be ruled out. Participants may have provided responses that aligned with their perception of the study's goals rather than their true reasoning processes.

The persistence of argument congruency bias, as evidenced in these studies, suggests that such biases are deeply ingrained and resistant to simple interventions. This resilience highlights

the potential necessity for more sustained and comprehensive strategies to effect meaningful changes in bias mitigation. These studies underscore the complexity of biases in argument evaluation and the challenges in effectively addressing these biases through straightforward cognitive interventions.

Overall, these findings point to the intricate nature of biases in argument evaluation and the substantial challenges in mitigating these biases through direct interventions. They emphasize the need for ongoing research into more effective bias reduction strategies, exploring deeper psychological mechanisms, and expanding these studies to include more diverse topics and populations to better understand and address the pervasive influence of biases in argument evaluation.

1.6 Conclusion

The first chapter of this dissertation, focusing on argument congruency bias and the potential of objectivity priming, contributes significantly to the larger narrative of how cognitive biases influence public discourse in today's polarized media landscape. By demonstrating how individuals preferentially evaluate arguments that align with their pre-existing beliefs, yet are capable of distinguishing between strong and weak reasoning, this chapter sheds light on the nuanced ways in which cognitive biases are manifested and can be countered. The exploration of objectivity priming as an intervention to mitigate these biases, although met with mixed results, provides valuable insights into the complexities of overcoming entrenched cognitive patterns.

The mixed results regarding the objectivity intervention suggest that the design and implementation of such interventions are crucial. Future research should explore the specific components that make these interventions effective, such as the inclusion of reflective writing

assignments or other engagement techniques. Additionally, investigating alternative methods to activate objectivity goals and examining their impact across diverse populations will be valuable.

Broader implications for future research include the need to develop and test interventions that can be scaled and applied in real-world settings, such as educational programs or media literacy campaigns. Understanding the mechanisms that reduce bias in argument evaluation can inform the design of tools and strategies to foster critical thinking and reduce the influence of misinformation. This work lays the foundation for future research to explore interventions that can successfully reduce bias in argument evaluation, ultimately contributing to a more informed and rational public discourse.

Overall, the chapter enhances our comprehension of the stubborn nature of cognitive biases and sets the stage for investigating broader strategies that might be more effective in promoting critical thinking and reducing the influence of misinformation in public discourse. This work is integral to the dissertation's broader aim of developing a comprehensive understanding of the factors that influence how individuals process information and form opinions in an increasingly complex and divided media environment.

Chapter 2

Correcting Misperceptions: What Role do Culturally-Relevant Interventions Play in Combating Misinformation?

The rapid proliferation of misinformation on social media, especially within political contexts, poses significant risks to public opinion and societal attitudes. This chapter examines the efficacy of corrective interventions in mitigating the influence of misinformation, with a particular focus on culturally-relevant sources. Leveraging the Elaboration Likelihood Model (ELM) and the concepts of hot and cold cognition, this research hypothesizes that corrections from in-group members, who share social or cultural identities with the recipients, will be more effective in reducing misperceptions than those from out-group members. Through two large-scale survey experiments (n=2,030 and n=1,502), targeting misinformation that affects Black and Latino communities, the studies reveal nuanced findings. While corrections generally reduce misperceptions, culturally-relevant corrections do not significantly outperform generic

corrections for Latino respondents and show limited advantages for Black respondents. The studies underscore the importance of source credibility and the context of misinformation, suggesting that the perceived source credibility is a crucial factor. This research contributes to the theoretical understanding of misinformation correction and offers practical insights for designing more effective interventions to combat misinformation in diverse communities.

2.1 Introduction

In an era where information spreads at unprecedented speeds, the digital landscape has become both a powerful tool for connection and a potential minefield of misinformation. As we navigate this complex terrain, the rapid proliferation of false or misleading content on social media platforms has emerged as a pressing concern, particularly within political spheres. This digital phenomenon has the potential to significantly sway public opinion and shape societal attitudes in ways that were unimaginable just a few decades ago (Allcott and Gentzkow 2017; Lazer et al. 2018; Vosoughi et al. 2018).

The impact of this misinformation is not merely abstract; it touches the lives of real people, influencing their decisions and behaviors in profound ways. To understand the human dimension of this issue, let's consider the story of Maria, a 32-year-old Latina living in a vibrant Los Angeles neighborhood.

Imagine Maria, scrolling through her social media feed in the months leading up to the 2020 U.S. presidential election. Like many, she uses these platforms to stay connected with friends and family, sharing life updates and engaging with her community. One day, a post from a close friend catches her eye. It warns that ICE agents are reportedly arresting people at local polling stations. The message sends a chill through Maria. Suddenly, the act of voting—a fundamental right she had always valued—becomes fraught with fear and

uncertainty. Concerned for her safety and that of her family, Maria seriously considers not voting.

However, Maria's story does not end there. A few days later, she encounters another post, this time shared by UnidosUS, an advocacy group she trusts. This post directly contradicts the earlier warning, clarifying that the information about ICE agents at polling stations is false. Relieved and reassured by this correction from a source she considers credible and culturally-relevant, Maria decides to cast her vote after all.

Maria's hypothetical experience vividly illustrates the potential power of misinformation to influence behavior, especially when it comes from trusted sources within one's community. While fictional, this scenario highlights an important dynamic: the potential for corrections, particularly those from culturally-relevant and trusted sources, to mitigate the impact of false information. Such examples, though not based on specific real-world events, are grounded in the broader patterns and concerns observed in research on misinformation and its effects on diverse communities.

This interplay between misinformation, corrections, and the sources of these messages forms the core of our investigation in this chapter. Building on the findings from Chapter 1, which explored the broader mechanisms of argument evaluation and bias, we now turn our attention to the specific challenge of combating misinformation in diverse communities.

Previous research has explored various interventions to counter misinformation, including fact-checking, media literacy campaigns, and the use of authoritative sources to provide corrections. However, the effectiveness of these interventions often depends on factors such as source credibility and message framing. The Elaboration Likelihood Model (ELM) and social identity theory offer valuable frameworks for understanding these dynamics. The ELM posits that persuasive communication can follow either a central or peripheral route,

influenced by factors such as the credibility of the source. Social identity theory suggests that individuals are more likely to accept information from sources they perceive as belonging to their in-group. This study builds on these frameworks by examining whether culturally-relevant corrections—those delivered by in-group members—are more effective at reducing misperceptions compared to corrections from out-group members.

This study explores the efficacy of corrective interventions aimed at mitigating the influence of misinformation, highlighting the impact of culturally-relevant sources in delivering corrections. Given that social media has transformed how information is disseminated, creating a landscape ripe for the quick spread of falsehoods, this research further explores whether corrections from in-group members—who share similar social or cultural identities with the recipients—are more effective at reducing misperceptions than those from out-group members, particularly in minority communities that are frequently the targets of misleading content (Tajfel et al. 1979; Reid 1987).

Misinformation's impact goes beyond individual beliefs, influencing attitudes and behaviors. The effectiveness of corrections varies, influenced by factors such as source credibility and message framing. Credible sources, particularly in-group members, are more likely to be trusted and effective in correcting misinformation (Ecker et al. 2014; Swire et al. 2017). Framing corrections in a way that aligns with the audience's values and beliefs can also enhance their persuasiveness (Nyhan and Reifler 2010).

Recent research in the fields of sociology, political science, and psychology has examined the dynamics of social categorization and its influence on individual behavior and attitudes. These studies emphasize how people categorize themselves and others into various social groups based on shared characteristics such as ethnicity, nationality, or political affiliation. This process of social categorization serves as a fundamental mechanism for individuals to

define their identity and establish a sense of belonging within their respective social contexts (Tajfel et al. 1979; Reid 1987).

Moreover, research suggests that these social categories not only shape individuals' perceptions of themselves but also influence their perceptions of others and their behavior towards them (Paluck et al. 2016). For example, individuals tend to exhibit favoritism when put under circumstances where they perceive competition or limited resources, leading them to prioritize the well-being of members within their own social group over others (Sherif et al. 1961). Similarly, in-group bias becomes apparent in situations where individuals identify strongly with a particular social group, leading them to show preferential treatment and allocate more resources to fellow in-group members compared to out-group members (Brewer 1999; Mullen et al. 1992).

On the other hand, prejudice and discrimination towards out-group members often manifest in various forms, such as unequal treatment in hiring practices (Pager 2007) or biased perceptions based on stereotypes and cultural differences (Devine 1989). These behaviors not only reflect individuals' social identities but also contribute to the perpetuation of inter-group conflicts and inequalities (Sidanius and Pratto 1996; Quinley 2009).

Given the influence of social identity, we hypothesize that corrections from in-group members will be more effective in reducing misperceptions about targeted misinformation. To test this, we conducted three experiments focusing on misinformation targeting Black and Latino communities. These studies used survey waves with oversampling of these populations to examine the effectiveness of in-group versus out-group corrections.

Study I demonstrated that corrective comments, in general, are effective in reducing misperceptions, but culturally-relevant corrections (corrective comments from an in-group member) did not show a significant advantage among Latino or Black respondents. Study II, which

focused on Black respondents, found that corrections from both Black and White candidates were effective, but there was no evidence that culturally-relevant corrections were more effective among Black participants.

Across both studies, in-group corrections were consistently the most effective, suggesting that social identity plays a crucial role in the acceptance of corrective information. Interestingly, the findings deviated from our initial expectations. We anticipated that individuals would be more receptive to corrections from members of their own group. However, the results indicated that when evaluating corrections about misinformation targeting a specific group, out-group members actually placed more trust in members of the targeted group for accurate information. This suggests a nuanced dynamic where the perceived expertise or credibility of the in-group source in relation to the specific misinformation plays a significant role in the effectiveness of the correction.

These results contribute to the understanding of how social identity influences the effectiveness of misinformation corrections. They highlight the potential of in-group corrections in reducing misperceptions, especially in contexts where misinformation targets specific ethnic groups. Interestingly, the findings challenge conventional notions of in-group bias and suggest a more complex dynamic where the perceived relevance and expertise of the source in relation to the misinformation play crucial roles.

2.2 Misinformation and Misperceptions

Misinformation refers to the dissemination of incorrect or misleading information and has become a pervasive issue in the digital age, particularly on social media platforms (Tandoc Jr et al. 2018; Lazer et al. 2018). The rapid spread of misinformation through these platforms can significantly impact public opinion and democratic processes, shaping political opinions,

influencing election outcomes, and undermining trust in democratic institutions (Allcott and Gentzkow 2017; Wardle and Derakhshan 2017). A notable example is the frequently circulated false claim during the 2020 U.S. presidential election that voting by mail leads to widespread voter fraud. This misinformation was propagated extensively on social media, creating doubts about the integrity of the electoral process and leading to significant public confusion and mistrust in the results (Benkler et al. 2020).

Misperceptions, or beliefs about facts that are either demonstrably false or unsupported by the best available evidence, can arise from misinformation (Flynn et al. 2017). These misperceptions can stem from various sources, including cognitive biases, erroneous inferences, misleading media coverage or even posts from trusted family friends (Flynn et al. 2017). Unlike mere ignorance, where an individual lacks information, misperceptions are often held with a high degree of certainty, leading individuals to consider themselves well-informed on a particular topic (Kuklinski and Quirk 2000; Nyhan and Reifler 2010). Persistent misperceptions, such as the denial of human-caused climate change, have hindered efforts to address environmental challenges despite overwhelming scientific consensus (Cook et al. 2016).

The persistence of misperceptions is often attributed to directionally-motivated reasoning, where individuals process information in a way that aligns with their pre-existing beliefs and attitudes (Taber and Lodge 2006b; Flynn et al. 2017). This cognitive bias leads to the selective acceptance of information that confirms one's beliefs and the dismissal of information that contradicts them, making it challenging to correct misperceptions once they are established (Nyhan and Reifler 2010). In the political arena, misperceptions can have far-reaching implications, influencing public opinion on important issues and shaping electoral outcomes (Flynn et al. 2017; Gaines et al. 2007).

Efforts to address misperceptions have taken various forms, from fact-checking initiatives to educational campaigns aimed at improving media literacy (Walter et al. 2020; Guess and Lerner 2020). However, the effectiveness of these interventions is mixed, with some studies suggesting that fact-checking can reduce misperceptions (Amazeen 2020), while others indicate that it may not be effective for deeply entrenched beliefs or in highly polarized environments (Nyhan et al. 2013).

2.2.1 Correcting Misperceptions

Preventive strategies for countering misinformation have become increasingly prominent in addressing the challenge of misleading information. Inoculation theory, initially proposed by McGuire in the context of persuasion, has been adapted to address misinformation in the political sphere (McGuire 1964). By exposing individuals to a weakened form of a misleading argument, this strategy aims to activate their cognitive defenses, making them less susceptible to misinformation when they encounter it in the future (Compton 2013; Banas and Rains 2010). Studies have demonstrated the effectiveness of inoculation in various political contexts, such as reducing the impact of negative political advertising and partisan attacks (Pfau et al. 2007; Compton et al. 2016).

Another critical pre-exposure strategy is media literacy education, which equips individuals with the skills to critically evaluate information, discern credible sources, and understand the mechanisms of media influence (Guess et al. 2020). Media literacy initiatives in the political domain focus on fostering a more discerning and informed electorate, thereby reducing the likelihood of individuals falling prey to misinformation. Research has shown that media literacy interventions can improve individuals' ability to recognize biases, assess the credibility of sources, and differentiate between factual information and misinformation (Jeong et al. 2009; Scheufele 2014). In the context of elections, media literacy programs have been developed to

educate voters about the tactics used in disinformation campaigns and the importance of verifying information before sharing it (Hobbs 1999; Kahne and Bowyer 2017).

Promoting analytical thinking aims to reduce the influence of misinformation by encouraging a more critical evaluation of information. This approach, rooted in the dual-process theory of cognition which posits two distinct modes of thinking, encourages individuals to activate their analytical system (Kahneman 2011). This aims to preemptively prompt a more thorough examination of information and its credibility, enhancing resilience against misinformation (Pennycook et al. 2021; Stanovich and West 2000). However, this strategy's success depends on individuals' motivation to think critically and their possession of the necessary cognitive resources (Jervis 2017).

The application of pre-exposure strategies in political science is not without challenges. Factors such as the individual's pre-existing beliefs, political ideology, and trust in media sources can influence the effectiveness of these interventions (Nyhan and Reifler 2010; Garrett et al. 2013). Additionally, the evolving nature of misinformation and the increasing sophistication of disinformation campaigns require continuous adaptation and innovation in pre-exposure strategies (Lewandowsky et al. 2017; Guess et al. 2019).

In the dynamic landscape of political information, individuals are inevitably exposed to misleading content despite preventive measures. This reality underscores the importance of post-exposure interventions, which focus on addressing misperceptions and correcting misinformation after individuals have already encountered it. These strategies are vital for mitigating the impact of misinformation that has already penetrated the cognitive defenses of the audience.

Post-exposure correction strategies

In the ongoing battle against misinformation, various post-exposure correction strategies have been developed to rectify false beliefs and promote accurate information. These approaches range from psychological interventions to direct factual corrections, each addressing different aspects of how people process and internalize information.

One effective approach is reframing, which involves constructing an alternative narrative that resonates with the audience's values and beliefs (Chong and Druckman 2007). This strategy is particularly potent in political contexts, where information is often interpreted through the lens of existing ideologies (Druckman 2010). For instance, reframing environmental issues in terms of economic benefits or national security can increase support among conservative audiences (Feinberg and Willer 2015).

Another promising strategy involves fostering analytical thinking. Interventions that encourage individuals to critically examine claims, such as asking them to explain how a statement could be true, have shown success in reducing belief in false information (Davis and Binning 2017). This approach helps engage the audience's critical faculties, making them less susceptible to misinformation.

These strategies acknowledge the challenge posed by motivated reasoning, where individuals selectively process information to reinforce preexisting beliefs (Ecker and Ang 2019; Hopkins et al. 2019). By addressing the psychological aspects of information processing, these approaches aim to create a more receptive environment for accurate information. Building on these foundational strategies, fact-checking has emerged as a cornerstone in the fight against misinformation. It serves as a crucial mechanism to directly clarify misconceptions

and promote informed decision-making. The effectiveness of fact-checking, however, can vary significantly depending on several factors.

Amazeen (2015) explored the impact of different correction formats, finding that contextual corrections—which provide explanations within the context of the claim—tend to be more effective than simple rating scales. These contextual corrections not only address the factual error but also educate the audience, potentially reducing the recurrence of similar misperceptions in the future.

The source of the correction plays a pivotal role in its effectiveness. Research by Nyhan et al. (2013) in the healthcare domain revealed that corrections from credible sources can reduce false beliefs. However, they also discovered a potential "backfire effect," where corrections can sometimes reinforce incorrect beliefs among those most ideologically committed to them. This underscores the importance of considering audience receptiveness and ideological alignment when delivering factual corrections.

In the realm of social media, the dynamics of fact-checking become even more complex. Margolin et al. (2018) found that on platforms like X, corrections are most effective when they come from users with high credibility and are directly linked to the misinformation. This highlights the importance of network structures and trusted sources in the dissemination and acceptance of corrections on social media platforms.

As misinformation becomes increasingly prevalent, especially in political discourse, developing more sophisticated and targeted fact-checking strategies is crucial. The effectiveness of these strategies depends on a nuanced understanding of the correction format, the credibility of the source, and the platform through which the correction is delivered. By considering these factors, along with the psychological insights gained from other correction strategies, we can work towards maintaining a well-informed public in the face of pervasive misinformation. This

research aims to contribute to this effort by examining the effectiveness of culturally-relevant fact-checking in combating misinformation in diverse communities.

2.3 Theory

This study's theoretical framework is grounded in cognitive processing models and social psychological theories of group dynamics (Petty and Cacioppo 1986; Tajfel 1974). We draw upon established concepts in information processing that describe how individuals evaluate and internalize new information, particularly in contexts where existing beliefs or identities may influence reception.

These cognitive models suggest that information can be processed through different mental pathways, each influenced by various factors including the complexity of the message, the receiver's motivation, and contextual cues (Taber and Lodge 2006b). Complementing these cognitive approaches, we also consider social psychological theories that explore how group affiliations and identities shape information reception and belief formation (Brewer 1979).

The interplay between these cognitive and social processes is particularly relevant in the context of misinformation and its correction. Our theoretical framework suggests that the effectiveness of corrective information may depend not only on its content but also on its source and its relationship to the receiver's social identity (Berinsky et al. 2017; Schuck et al. 2017).

By integrating these perspectives, we develop a nuanced approach to understanding how individuals might respond to corrections of misinformation, especially when these corrections come from sources with varying degrees of cultural or group relevance. This theoretical

foundation informs our hypotheses about the general efficacy of corrections and the potential enhanced impact of culturally aligned sources in mitigating misinformation beliefs.

2.3.1 Hypotheses

Research has demonstrated that exposure to corrective information can effectively reduce misperceptions across various contexts. Studies such as those by Nyhan and Reifler (2010) and Thorson (2016) have shown that presenting individuals with information that directly challenges their misperceptions, whether about political issues or misinformation encountered on social media, can lead to a significant update in their beliefs. Additionally, Vraga and Tully (2020) highlighted the role of news literacy in enhancing individuals' ability to engage with and accept corrective information, emphasizing the need for both access to accurate information and the development of critical evaluation skills to combat misinformation effectively. These studies collectively underscore the potential of exposure to corrective information in reducing misperceptions. By directly challenging false beliefs with factual information, it is possible to encourage individuals to update their beliefs and reduce the influence of misinformation.

Hypothesis 1: Corrective comments will reduce misinformation belief relative to the no correction and control conditions.

To understand our expectations for culturally-relevant corrections on social media, it's helpful to start with the Elaboration Likelihood Model (ELM), which outlines two pathways for processing information: the central route and the peripheral route (Petty and Cacioppo 1986). The central route involves careful scrutiny of the content, while the peripheral route relies on heuristics or shortcuts, such as the credibility of the source or the number of likes on a post. The ELM has been widely applied in various contexts, including political communication

and persuasion, to understand how different factors influence the processing of persuasive messages (Petty and Cacioppo 1986; Taber and Lodge 2006b).

The concepts of hot and cold cognition in political science relate closely to the ELM's pathways. Hot cognition refers to emotionally charged, reflexive processing of information, often influenced by pre-existing beliefs and biases, akin to the peripheral route. Cold cognition, on the other hand, involves a more analytical and deliberate evaluation of information, similar to the central route (Taber and Lodge 2006b). Research in political psychology has further explored these concepts, demonstrating how emotional responses and cognitive biases can impact political judgment and decision-making (Redlawsk 2002).

Heuristics play a significant role in hot cognition, where individuals use shortcuts to quickly assess information. In the context of political misinformation, one important heuristic is in-group membership. Research has shown that individuals are more likely to trust and accept information from sources that are perceived as part of their in-group, as it aligns with their identity and values (Berinsky et al. 2017; Schuck et al. 2017). Studies on in-group bias have demonstrated how this heuristic can influence attitudes and behaviors, leading to preferential treatment of in-group members and discrimination against out-group members (Tajfel 1974).

Hypothesis 2: Corrections from culturally relevant organizations will more effectively reduce misinformation belief

Research in psychology and sociology has extensively explored the impact of in-group bias on various aspects of human behavior. In-group bias leads individuals to favor members of their own group over those of out-groups, affecting everything from interpersonal interactions to decision-making processes (Brewer 1979). This bias can manifest in various forms, such as

preferential treatment, positive evaluations, and increased trust and cooperation within the in-group (Hewstone et al. 2002).

In the political domain, in-group bias can significantly shape individuals' responses to information and misinformation. Studies have shown that people are more likely to believe and spread information that aligns with their in-group's beliefs and to dismiss or ignore information that contradicts these beliefs (Greene 2004). This can lead to the reinforcement of existing narratives and the polarization of opinions, as individuals become more entrenched in their in-group's viewpoint (Iyengar et al. 2019).

The influence of in-group bias on information processing is particularly relevant when considering the dynamics of hot cognition in the context of political misinformation. We posit that when individuals encounter corrective information from an in-group source, the emotional and identity-driven aspects of hot cognition are likely to come into play, leading to a higher perceived credibility and trustworthiness of the information. This increased trust can facilitate the acceptance and updating of beliefs, as the information aligns with the individual's in-group identity and/or values (Vraga et al. 2022).

To address these questions, we conducted two large-scale, nationally representative survey experiments, targeting misinformation prevalent during the 2020 election cycle. The entire experiment, including these two hypotheses, was pre-regisitred with the Evidence in Governance and Politics (EGAP), to promote rigourous research. The entire pre-registration document can be found on page 143 of Appendix B.

The experiments employed a randomized controlled design, with participants assigned to one of several treatment conditions or a control group. The treatment conditions varied in terms of the source of the correction, with some corrections provided by a culturally-relevant source (in-group member) and others by a generic source. A baseline group was exposed to

the misinformation without any correction and a control group did not see the Facebook post at all. This design allowed us to isolate the effect of the source's cultural relevance on the effectiveness of the correction.

In both studies, we measured participants' misperceptions before and after exposure to the treatments. This pre-/post- measurement approach enabled us to assess the impact of the corrections on participants' beliefs. The primary dependent variable was the change in misperceptions, operationalized as the difference between pre- and post-treatment scores.

The initial study consisted of two distinct experiments: one involving Latino participants and another involving Black participants. Both experiments targeted misinformation suggesting that ICE agents and local police were apprehending individuals at polling stations. The subsequent study, however, focused exclusively on Black participants and examined misinformation related to voter suppression.

While the overarching research design was consistent across both studies, there were some differences in the specifics of the treatment conditions and the misinformation targeted. In Study 1, the culturally-relevant source was UnidosUS for Latino participants and the National Association for the Advancement of Colored People (NAACP) for Black participants. In Study 2, the focus was limited to Black participants, and the culturally-relevant source was a Black congressional candidate. These variations in the treatment conditions and target misinformation were tailored to the specific context and demographic focus of each study, allowing for a nuanced exploration of the effectiveness of culturally-relevant corrections in different settings.

The study was approved by the Washington University in St. Louis (WashU) Institutional Review Board (IRB), ensuring that all research activities adhered to ethical guidelines for the protection of participants. Given the sensitive nature of exposing participants to

misinformation, the study design incorporated a critical measure to address ethical concerns: debriefing sessions after the experiments. In these sessions, participants were informed that the misinformation they encountered was intentionally false and used solely for research purposes, ensuring they left the study with accurate information.

2.4 Study I

Study I comprised two separate experiments, each tailored to assess the impact of targeted misinformation on voter behavior, with a focus on the efficacy of corrective feedback. Experiment 1 centered on misinformation aimed at Latino voters, suggesting ICE agents were detaining people at polling stations. Experiment 2 examined a parallel scenario targeting Black voters, with misinformation about local police apprehending individuals with outstanding warrants at polling sites.

2.4.1 Research Design

The chosen misinformation topics were purposefully selected for their potential to suppress voter turnout, a key concern in political science. These topics reflect strategies used by foreign actors throughout the 2020 election, where known falsehoods (disinformation) were spread by unsuspecting users (misinformation).[1] It is believed that these tactics aim to reduce election participation and erode trust in democracy by perpetuating false beliefs about voting requirements, disseminating misleading content about the consequences of voting, and using social media platforms to amplify divisive narratives (The Brennan Center 2022).

[1]Disinformation is defined as intentionally false or misleading information, while misinformation refers to information that is false or inaccurate but not created with the intention of causing harm (Aïmeur et al. 2023).

Variables

Our primary dependent variable for this study was *misperceptions*, measured on a 5-point Likert scale, with higher values indicating stronger beliefs in the misinformation. Prior to exposing participants to a treatment condition, we asked them to report their beliefs regarding the accuracy of the claims that immigration officials (in Experiment 1) and local police (in Experiment 2) were arresting people at the polls on election day in 2020.[2]

Table 2.1: Summary of treatment groups

Treatment group	Description
No-correction	Participants saw a post with misinformation only.
Generic source correction	Correction provided by the ACLU.
Culturally-relevant correction	Correction from UnidosUS or the NAACP.[3]
Control	No social media posts were viewed.

The exact treatment conditions can be reviewed in page 147 and page 148 of Appendix B.

After completing other unrelated survey questions, participants were then exposed to the treatment conditions and subsequently asked about their beliefs in the misinformation.[4] This pre-/post-measurement of misperceptions allows us to assess the impact of the corrections using a within-subject design to understand the extent to which misinformation can increase beliefs in misperceptions.

[2]See page 146 for exact language of the misperceptions measure.

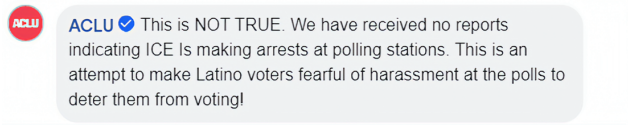
[3]UnidosUS was the culturally-relevant source in Experiment 1 and the NAACP was the culturally-relevant source in Experiment 2.

[4]Randomization for each experiment was independent (i.e. participants could be in the control condition in Experiment 1 and the no correction condition in Experiment 2). The order in which the two experiments were presented to participants was also randomized.

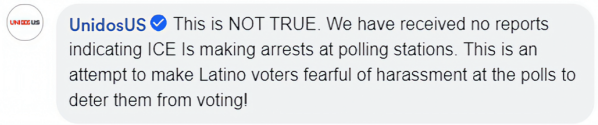
Table 2.1 provides an overview of our treatment conditions which serves as our main independent variables throughout both analyses. This between-subject comparison among treatment groups enabled an experimental examination of the effects of corrections.

Figure 2.1: Experiment 1 treatment conditions

(a) Facebook post with no correction



(b) Generic reply



(c) culturally-relevant reply

The control condition saw neither the post nor the correction and still received the pre-/post-treatment beliefs in these misperceptions

This comparison is critical to understanding the role of source and fact-check interventions in the correction of misinformation. By using Facebook post templates, the study taps into

the routine engagement with social media, thus enhancing the real-world applicability of the results.

2.4.2 Results

Experiment 1: Latino respondents

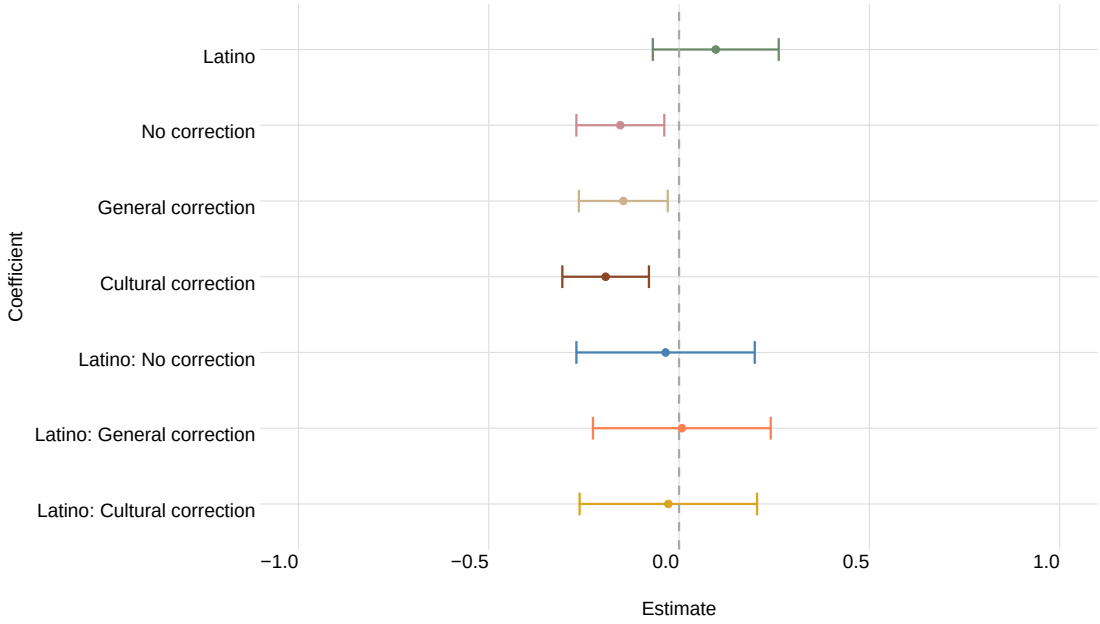
The results from the first experiment, which focused on Latino participants, are visualized in Figure 2.2. Misperceptions were assessed using a 5-point Likert scale, where higher values indicate more belief in the false claims. The baseline for comparison was participants who only encountered the post with misinformation.

The figure shows the effectiveness of different correction strategies in reducing misperceptions. The first coefficient displays the impact of being Latino on beliefs in misperceptions, overall. Being Latino slightly increases one's likelihood of believing in the misinformation claiming ICE is targeting election polls, however, this increase in belief in misperceptions is neither substantial nor significant.

Moving on to the treatment condition coefficient, the culturally-relevant correction ($\beta = -0.19$, $se = 0.06$, $p < 0.001$) had the most impact, however the values among those who saw no correction and those who saw the general correction were similar.

The specific breakdown for Latino participants shows a similar trend. When corrections were culturally tailored, there was a notable decrease in misperceptions among Latino participants, although this reduction was not statistically significant compared to the overall sample. This pattern confirms that while culturally-relevant corrections resonate well, their unique effect on Latino participants compared to the broader group requires further exploration.

Figure 2.2: Effect of exposure to culturally-relevant correction on misperceptions among Latino participants



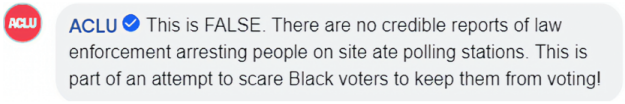
Illustrates the regression analysis of the impact of different correction strategies on the misperceptions held by Latino participants regarding misinformation. The x-axis represents the regression coefficients, and the y-axis denotes the various correction strategies: culturally relevant correction, general correction, and no correction, applied to Latino respondents. The 95% confidence intervals are shown, which help gauge the precision of the estimated effects. Negative coefficients suggest a reduction in misperceptions, indicating the effectiveness of the correction strategy. See page 149 in Appendix B for detailed regression results.

Experiment 2: Black respondents

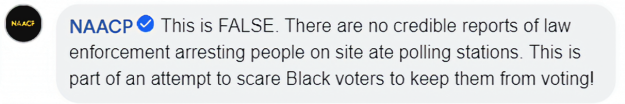
In the second experiment, we replicated the design with Black participants, focusing on misinformation about local police arresting individuals at polling stations. The results, depicted in Figure 2.4, show the regression coefficients for various correction strategies. The first coefficient displays the impact of being Black on beliefs in misperceptions, overall. Being Black increases one's likelihood of believing the misperceptions posited by the misinformation. This heightened susceptibility is understandable given that the misinformation specifically

Figure 2.3: Experiment 2 treatment conditions

(a) Facebook post with no correction



(b) Generic reply

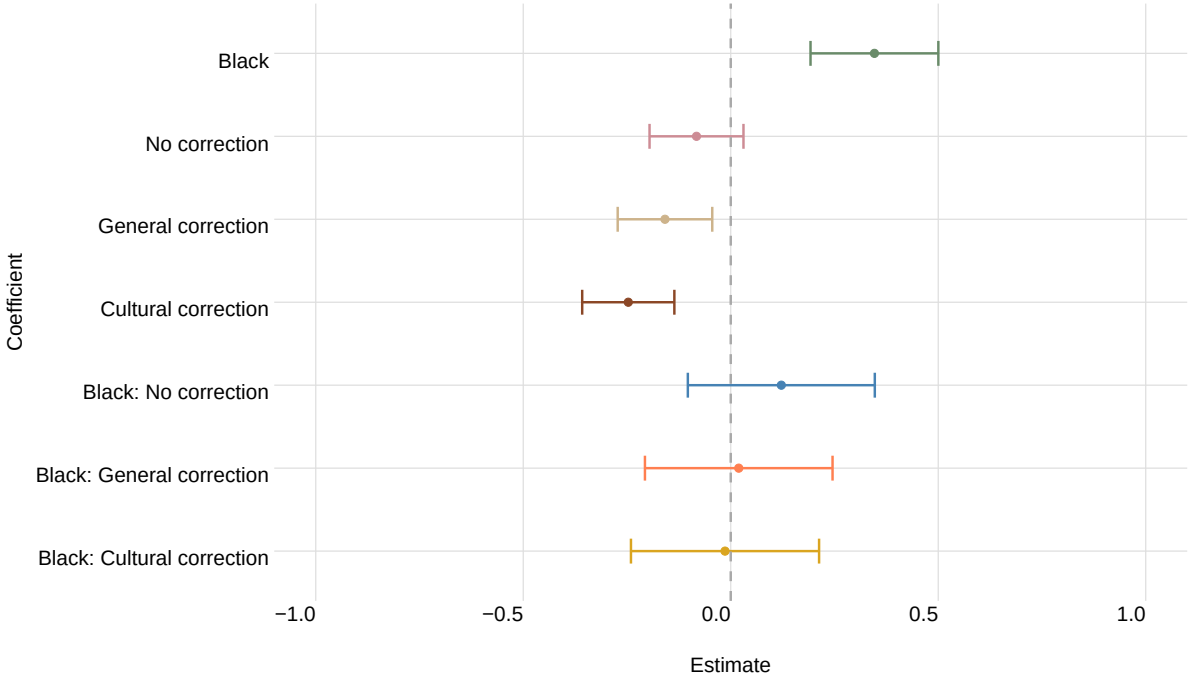


(c) Culturally-relevant reply

targets the Black community, potentially making the message more salient and concerning to Black participants.

The data illustrate that while all types of corrections reduce misperceptions, the culturally-relevant corrections (marked by more substantial negative coefficients) are notably more effective. These results confirm the utility of targeted misinformation corrections among Black populations in reducing erroneous beliefs about police activities at polling stations. Interestingly, the impact on Black participants does not significantly differ from the overall sample.

Figure 2.4: Effect of exposure to culturally-relevant correction on misperceptions among Black participants



Represents a regression analysis that quantifies the effect of various correction strategies on misperceptions among Black participants, specifically in the context of misinformation related to local police activities at polling stations. The x-axis displays the regression coefficients, while the y-axis lists the groups and correction strategies tested: no correction, general correction, and culturally-relevant correction both within and outside the Black community. The 95% confidence intervals are shown for each coefficient, providing a measure of the estimate's uncertainty. See page 150 of Appendix B for regressions.

These findings carry significant implications for countering misinformation on social media, particularly during elections, underscoring that while culturally-tailored interventions are beneficial, their specific efficacy on targeted demographic groups like Black participants may require further research to optimize impact.

The findings from both experiments suggest that corrections to misinformation on social media can effectively reduce misperceptions among the general population.[5] The culturally-relevant corrections, provided by organizations like UnidosUS and the NAACP, appear to be particularly effective in reducing misperceptions. However, contrary to our expectations, these corrections did not have a significantly greater impact on the targeted demographic groups (Latino and Black participants) compared to the overall sample.

These results have important implications for combating misinformation on social media, especially in the context of elections. They suggest that corrections, particularly those from culturally-relevant sources, can be an effective tool in reducing the impact of misinformation. However, the lack of a significantly greater impact on the targeted groups indicates that further research is needed to understand how to tailor corrections more effectively to specific demographic groups.

2.5 Study II

Study II builds upon the findings of Study I by investigating how the race of the source providing corrections affects their effectiveness in reducing misperceptions about voter suppression. This study specifically examines whether corrections from (hypothetical) Black or White congressional candidates have differing impacts on the beliefs of Black participants. By focusing on the racial identity of the correction source, this research aims to further our understanding of how culturally-relevant factors influence the acceptance of corrective information.

[5] Additional regressions examining both experiments among White participants only can be found on page 151 of Appendix B.

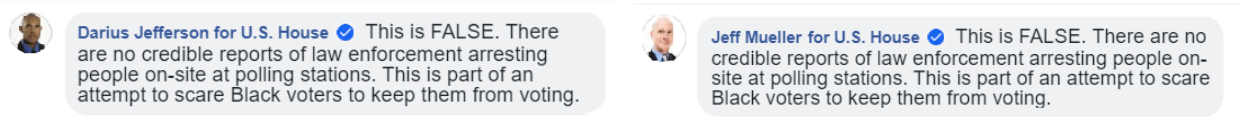
2.5.1 Research Design

The design mirrors that of Study I, employing a randomized controlled approach to assess the effectiveness of culturally-relevant corrections compared to generic ones. By building on the research design and findings of Study I, Study II aimed to provide a deeper understanding of the dynamics of misinformation correction and the role of culturally-relevant corrections in influencing beliefs within specific demographic groups.

Variables

The primary independent variable, similar to Study I, is the treatment condition assigned to participants. The treatment conditions in Study II, detailed in Figure 2.5, differ slightly from those in Study I.[6] The rationale behind these treatment groups remains consistent with the first study, aiming to determine whether the race of the corrective source affects its credibility and, consequently, its effectiveness in mitigating misperceptions. This is particularly pertinent in the context of voter suppression, a concern that disproportionately impacts minority communities.

Figure 2.5: Study II treatment conditions



(a) Black candidate correction (b) White candidate correction

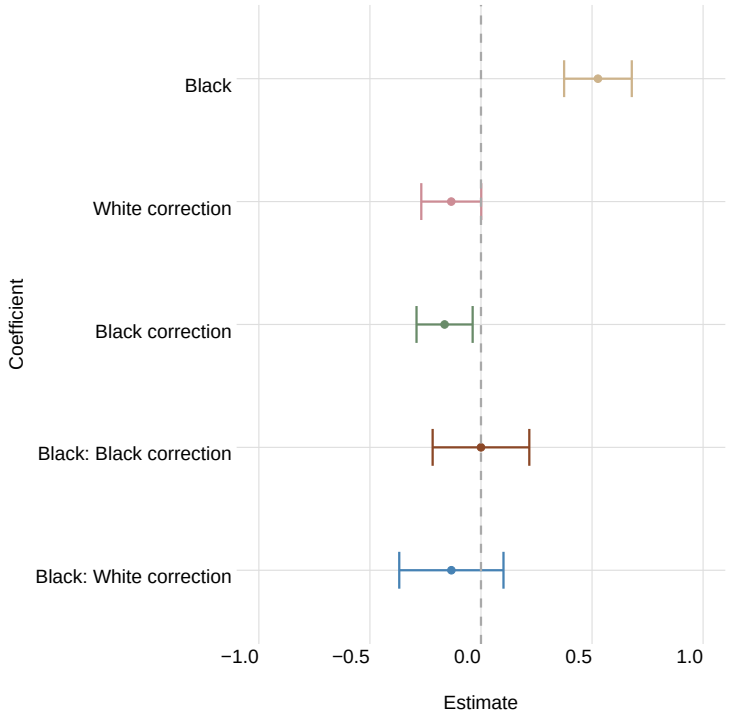
Study II relied on the same misinformation post as Study I, displayed in Figure 2.3a.

[6]Study II exposed all participants to the Facebook post.

The primary dependent variable, as in Study I, is the level of misperceptions held by participants, measured before and after exposure to the experimental treatments. Misperceptions are assessed using a statement related to voter suppression, with participants rating their confidence in its accuracy on a 4-point Likert scale, with higher values indicating more belief in the misinfomration.

2.5.2 Results

Figure 2.6: Effect of exposure to culturally-relevant correction on misperceptions among Black participants



Presents the results of a regression analysis examining the impact of different correction types on misperceptions among entire sample and also an interaction displaying Black participants only. Displaying coefficients and their 95% confidence intervals. The horizontal axis represents the estimated effect size, while the vertical axis lists the different correction conditions. See page 154 of Appendix B for associated regression analysis.

In Study II, we assess the impact of corrections on misperceptions related to voter suppression among Black participants. Figure 2.6 illustrates the effects of different corrections on misperceptions among the entire sample and among Black participants in particular. Like Study I, Black participants, regardless of condition, were more susceptible to belief in the misinformation.

Examining the post-treatment scores for misperceptions, we observed a notable effect for the correction provided by the Black candidate correction ($\beta = -0.18$, $se = 0.07$, $p < 0.01$). and the White candidate correction($\beta = -0.13$, $se = 0.07$, $p = 0.09$), with the latter being slightly less effective. Observing these corrections among Black participants only revealed that the culturally-relevant correction was *not* more effective.

2.5.3 Discussion

The findings from Study I and Study II provide nuanced insights into the effectiveness of culturally-relevant corrections in addressing misinformation among minority communities. In Study I, which focused on Latino and Black participants, we observed that while corrections in general were effective at reducing misperceptions, culturally-relevant corrections did not show a significant advantage among Latino respondents. Among Black participants, both culturally-relevant and general corrections were effective, but the anticipated superiority of culturally-relevant corrections was not evident.

Study II extended the investigation by focusing on Black participants and the role of the race of the corrective source. The results indicated that corrections from both Black and White candidates were effective in reducing misperceptions. Interestingly, there was no significant interaction effect, suggesting that the race of the respondent did not significantly influence

the effectiveness of the corrections. This finding challenges the hypothesis that in-group corrections (i.e., corrections from a member of the same racial group) would be more effective.

One possible explanation for these results is the nature of the misinformation itself. Voter suppression is a topic that disproportionately affects Black communities, and corrections from credible sources, regardless of racial congruence, may be equally persuasive. The perceived expertise or relevance of the source in relation to the specific misinformation might play a more crucial role than in-group identity. For example, corrections from a Black candidate on an issue like voter suppression might be generally more persuasive due to the candidate's perceived credibility and authority on the subject, rather than solely because of shared racial identity.

2.6 Conclusion

The findings from these studies make significant theoretical contributions to the literature on misinformation and social identity. Contrary to the conventional notion of in-group bias, our results suggest a more complex dynamic where the perceived relevance and expertise of the source in relation to the misinformation play crucial roles. While culturally-relevant corrections are beneficial, their specific efficacy on targeted demographic groups like Latino and Black participants may not be as straightforward as initially hypothesized.

The broader implications of these findings are substantial for the design of interventions aimed at combating misinformation. For practitioners and policymakers, the results underscore the importance of source credibility and the need to consider the specific context and nature of the misinformation when designing corrective messages. While in-group sources can be effective, out-group sources with high perceived relevance and expertise can also play a crucial role in reducing misperceptions.

Future research should continue to explore the nuances of source credibility and the conditions under which culturally-relevant corrections are most effective. Understanding the interplay between social identity, source credibility, and message framing will be essential for developing more effective strategies to counteract misinformation and its detrimental effects on public opinion and democratic processes.

Chapter 2 has built upon and extended the insights gained from Chapter 1's exploration of argument congruency bias and objectivity priming. While Chapter 1 demonstrated that individuals can distinguish between strong and weak arguments but are influenced by their pre-existing beliefs, Chapter 2 delves deeper into how these biases operate in the specific context of misinformation correction among minority communities.

The results of this study provide a fascinating counterpoint to the argument congruency bias observed in Chapter 1. While we might expect culturally-relevant corrections to be more effective due to in-group bias, our findings suggest a more nuanced picture. The effectiveness of corrections seems to depend more on the perceived expertise and relevance of the source in relation to the specific misinformation, rather than solely on cultural or racial congruence. This highlights the complex interplay between cognitive biases, social identity, and the specific content of information being processed.

Furthermore, these findings bridge the gap between the theoretical concepts explored in Chapter 1 and the real-world application of correction strategies examined in Chapter 3. They suggest that while cognitive biases and social identity play important roles in information processing, the effectiveness of interventions to combat misinformation may depend on a broader range of factors than initially hypothesized. This underscores the need for a multifaceted approach to addressing misinformation, one that considers not only the cognitive

and social psychological aspects of information processing but also the specific context and content of the misinformation itself.

Chapter 3

From Posts to Perceptions: Can Racially-Charged Social Media Content Impact Attitudes and Opinions?

This chapter investigates the influence of social media interactions on the racial attitudes of White Americans, particularly focusing on their reactions to counter-normative speech concerning race. Utilizing an experimental framework, the study engaged 3,500 participants from Amazon's Mechanical Turk to respond to a controlled set of social media posts and their varied replies, which included endorsements of racial slurs, condemnations, and mixed responses. The primary goal was to assess how exposure to these differing viewpoints impacts individual opinions and broader racial attitudes. Findings indicate that despite the dynamic and interactive nature of social media, changes in racial perceptions are minimal, suggesting deeply entrenched beliefs that are resistant to immediate influence through digital discourse. This chapter contributes to the understanding of social media's role in political communication

and highlights the complexities of using digital platforms as tools for social change in racial attitudes.

3.1 Introduction

The trajectory of racial discourse in the United States has undergone significant transformations, deeply embedded in the fabric of the nation's political and social landscapes. Historically, race was used for political maneuvering and suppression of civil rights, especially during the Jim Crow era in the South (Allport 1954). The civil rights movement challenged overtly racist dialogues, leading to subtler forms of racial communication (Pettigrew and Tropp 2006). This shift was characterized by "dog whistle" politics, where coded language addressed racial issues without direct backlash (Mendelberg 2001; Lopez 2016).

The election of Barack Obama marked another pivotal moment in racial discourse, catalyzing explicit racial discussions, such as the birther movement, which revealed persistent undercurrents of prejudice (Binder et al. 2009). The advent of social media further transformed racial rhetoric, amplifying explicit racial discourse and intertwining historical racial tensions with modern digital interactions (Sunstein 2018). This evolution underscores the need to understand these dynamics to address challenges in political communication and racial discourse in the digital age (Plant and Devine 1998; Blanchard et al. 1994).

The rise of social media has intensified these dynamics, facilitating the spread of diverse viewpoints and amplifying extreme, racially charged content (Sunstein 2018; Tucker et al. 2018). Social media platforms have democratized access to information, allowing users from all backgrounds to share their perspectives. This inclusivity has enabled significant social movements to gain traction and mobilize support on an unprecedented scale. However,

alongside this democratization, social media also provides a venue for the spread of divisive and extremist views, which can polarize public discourse and exacerbate social tensions.

Algorithmic biases on social media platforms play a crucial role in this polarization. These algorithms are designed to maximize user engagement by promoting content that elicits strong emotional reactions (Allcott and Gentzkow 2017; Bail et al. 2018; Guess et al. 2018). This often means that sensationalist and emotionally charged posts are more likely to be seen and shared, regardless of their factual accuracy or social impact. As a result, users are frequently exposed to content that reinforces their preexisting beliefs and biases, creating a feedback loop that deepens ideological divides (Crockett 2017). This curation fosters environments where extreme views can thrive, and moderate voices are often drowned out.

This research aims to examine how social media influences racial discourse, particularly focusing on its role in the resurgence of explicit racial expressions. This study hypothesizes that social media serves as a powerful medium for the dissemination and amplification of racial ideologies, both normative and counter-normative. The hypothesis posits that engagement with counter-normative racial comments on social media platforms leads to an increase in negative racial attitudes among users and more favorable views of the announcer compared to tweets without counter-normative comments.

To test this hypothesis, the research design involves exposing participants to various types of social media comments about a racially charged incident and assessing their subsequent attitudes. The study's findings reveal a nuanced understanding of how social media interactions influence racial attitudes, indicating a notable resilience in these attitudes despite varied exposures . This suggests that deeply ingrained societal views and biases are not easily altered by brief social media interactions.

The implications of this research are significant for understanding the role of digital platforms in racial discourse. The findings highlight the complexity of social media's influence on racial attitudes, suggesting that while these platforms provide a space for diverse expressions, they may not significantly shift deeply held beliefs in the short term. This study contributes to the broader literature by providing insights into the intricate relationship between digital media consumption and the persistence of racial attitudes in contemporary society. It raises important questions about the long-term effects of repeated exposure to diverse racial narratives on social media and the potential for these platforms to either reinforce or gradually shift societal norms regarding race.

3.2 Racial Norms in Contemporary America

The trajectory of racial discourse in the United States has evolved continuously, deeply intertwined with the nation's political and social fabric. Historically, race has been used by politicians to secure power and suppress civil rights, particularly during the Jim Crow era (Allport 1954). As society's awareness of racial injustices grew, the civil rights movement marked a shift away from overtly racist dialogues (Pettigrew and Tropp 2006).

In the late 20th and early 21st centuries, "dog whistle" politics emerged, using coded language to address racial issues without overt racism (Mendelberg 2001; Lopez 2016). Barack Obama's election marked another shift, with his presidency catalyzing a resurgence of explicit racial discussions, despite advancements in racial equality (Binder et al. 2009).

Social media has amplified explicit racial discourse, merging historical racial tensions with modern digital interaction, highlighting a continuum from blatant exploitation to coded messages and back to overt expressions (Sunstein 2018). Understanding these influences is vital for addressing the challenges posed by the digital age in political communication.

3.2.1 Evolution of Racial Norms in Political Discourse

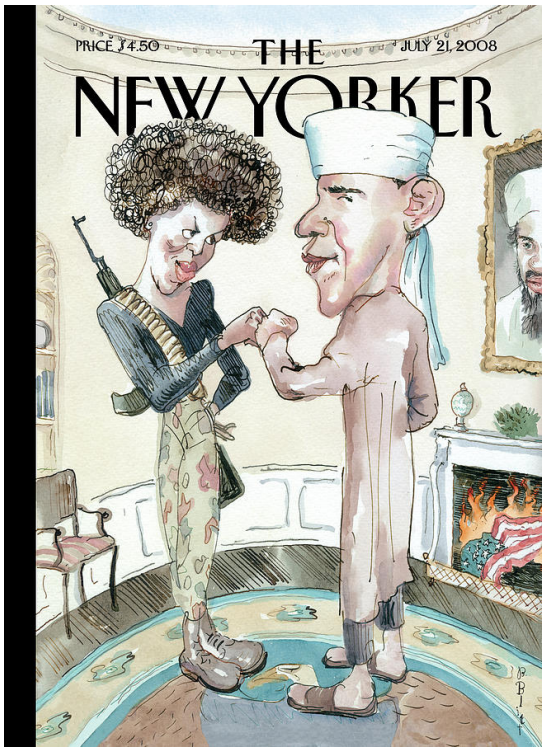
Since the early 21^{st} century, the rejection of explicit racism in the United States has weakened, influenced by shifting societal norms and political rhetoric (Hutchings et al. 2010; Piston 2010; Huddy and Feldman 2009; Valentino et al. 2018; Reny et al. 2020). Historically, overt racial cues were replaced by subtler messages that resonated without evoking social stigmas, as explored in Mendelberg (2001) *The Race Card*. Explicit racial appeals became less effective because they highlighted a societal commitment to racial equality and discomfort with violating this norm (Gilens 1996; Valentino et al. 2002).

Post-civil rights era, American society increasingly disapproved of 'old-school' racism, which directly attributes disparities faced by Black Americans to inherent biological deficiencies. Instead, a 'new racism' emerged, grounded in cultural rather than biological critiques. This new form of racism—encompassing beliefs such as 'laissez-faire racism,' 'aversive racism,' and 'symbolic racism'—suggests that the challenges faced by Black communities are a result of cultural deficiencies, like a lack of work ethic or over-reliance on welfare (Bobo and Kluegel 1998; Dovidio and Gaertner 2004; Sears 1988; Sniderman et al. 1991).

Political scientists adapted methodologies to capture subtler forms of modern racism, employing implicit racial cues like 'inner city' and 'tough on crime,' reflecting coded language used by politicians to signal racial biases without overtly violating racial equality norms (Mendelberg 2001). These changes illustrate a strategic evolution in political communication, aligning with nuanced public sentiments regarding race while still activating underlying racial attitudes (White 2007; Banks and Valentino 2012; Valentino et al. 2002).

3.2.2 The "Post-Racial" Era and Apparent Shifts in Explicit Appeals

Figure 3.1: "The Politics of Fear" by Barry Blit



Barack Obama's presidency marked a shift in racial discourse, with more explicit racial dialogue emerging. The birther movement, led by figures like Donald Trump, questioned Obama's legitimacy based on his racial background (Tesler 2012; Serwer 2019).

The Tea Party movement used explicitly racist imagery, like depictions of Obama as an African witch doctor and circulated emails portraying him as an ape (Los Angeles Times 2010; CNN Wire Staff 2011). Such acts embraced overt racism to energize voter segments.

Mainstream media mirrored these sentiments under the pretense of satire. The New Yorker's cover featuring the Obamas as militants aimed to criticize stereotypes but blurred lines between satire and reinforcement of stereotypes (Halttunen 2008). Fox News's reference to

Michelle Obama as a "baby mama" and Rush Limbaugh's "Barack the Magic Negro" employed racial caricature to mock the president (Parks 2010; Limbaugh 2007).

These instances in the media were indicative of a broader shift in the political landscape. They demonstrated a departure from the covert racial appeals described by Mendelberg (2001), as they were overtly racist in nature. The willingness of prominent media figures and outlets to engage in such explicit racial discourse suggested a changing tolerance for racism in public discourse. This trend not only challenged the existing theories of racial priming but also raised questions about the role of the media in shaping and reflecting societal attitudes towards race. These public instances of explicit and charged racial cues were not isolated but rather represented a number of pressures making identity issues ever salient and often framed problematically.

3.2.3 Pressures Toward Racial Animus among White Americans

The influence of race in American politics has historically been profound, with emotions such as anger and fear frequently underpinning prejudices held by White Americans towards racial minorities (Banks 2014; Stephan et al. 2002). This phenomenon has been further compounded by events and societal shifts that have intensified racial identities and prejudices. It is simplistic to attribute the rise in overt racist expressions solely to the election of President Obama; rather, a convergence of factors around his election brought racial identity to the forefront of political discourse.

The notion of a White majority facing a status quo threat has made racial identity "chronically salient" for many White Americans, guiding political judgments and behaviors (Jardina 2019). A strong in-group identity among White Americans can lead to more explicit expressions of out-group prejudice (Effron and Knowles 2015).

Changing racial demographics in the United States play a crucial role. The projection that White Americans are losing their majority status contributes to a fear of losing political power, reflected in the Democratic members' racial composition in the House of Representatives (Poston and Sáenz 2020; Newport 2018).

The Black Lives Matter and #MeToo movements highlight racial inequalities and advocate for inclusivity, with 'cancel culture' emerging to ostracize individuals exhibiting racist or sexist behavior, though perceptions vary between political affiliations (Vogels et al. 2021).

If perceptions of rising explicit racism are correct, it suggests factors like the election of a non-White president, demographic shifts towards a diverse society, and cultural movements advocating for racial justice are perceived as threats, enhancing in-group solidarity and out-group animosity among White Americans (Vogels et al. 2021; Doosje et al. 2002; Grant and Brown 1995; Kinder and Kam 2010; Lau 1989). The changing societal landscape challenges traditional White privileges as America moves towards a more inclusive identity.

The shift from overt racial discourse to nuanced expressions and back sets the stage for understanding the digital platforms' impact today. The reemergence of explicit racial expressions in politics and media requires exploring how these dynamics are complicated and amplified by digital revolution. The internet and social media have become arenas where racial issues are discussed and intensified, challenging traditional racial discourse boundaries. This intersection of historical racial tensions and modern digital communication necessitates examining how racial norms are contested and reshaped in the digital age, highlighting digital platforms' role in racial discourse.

3.3 Racial Discourse in the Digital Age

The legacy of racial dynamics in politics, deeply influenced by America's racial history, continues to shape modern political discourse. As technology evolves, scholars examine how digital platforms impact racial polarization, emphasizing the importance of this research in understanding the influence of diverse online perspectives on personal attitudes (Cobb 2016; Valentino et al. 2018; Tesler 2018; Noble 2018; López 2014).

Social media democratizes access to varied viewpoints, enabling engagement with diverse political and racial discussions. However, it also facilitates the spread of extreme views, potentially leading to more pronounced ideological divides (Sunstein 2018; Tucker et al. 2018). The curated nature of social media feeds can reinforce these divides (Pariser 2011; Allcott and Gentzkow 2017), though some studies suggest exposure to broader viewpoints could mitigate polarization (Barberá et al. 2015).

The role of social media in bypassing traditional gatekeepers allows unchecked dissemination of content, often amplifying harmful narratives and impacting racial discourse significantly (Pew Research Center 2020; Allcott and Gentzkow 2017; Mutz 2018). Anonymity and the lack of accountability on these platforms can increase hostile behaviors, which underscores the need for a deeper understanding of social media's role in societal divisions (Cheung et al. 2021; Goswami 2018; Sunstein 2018; Gillespie 2018).

3.3.1 Social media: The Good, the Bad, and the Ugly

The examination of social media's impact on racial discourse reveals a layered influence that spans positive mobilization and educational outreach to negative polarization and the proliferation of hate speech. This section delineates these dynamics to underscore how

social media serves as both a catalyst for progressive social change and a platform that may perpetuate division and intolerance. The interaction of these influences with past and present racial norms offers a comprehensive perspective on the evolving narrative of racial dialogue facilitated by digital platforms.

The Good: Uplifting underrepresented voices, amplifying activism, and cultivating connections

Social media platforms like X and Facebook have become critical for highlighting voices that are traditionally marginalized in mainstream discourse. Movements such as Black Lives Matter have effectively used these platforms to spotlight racial injustices, showcasing social media's pivotal role in broadcasting diverse narratives and fostering a global community of support and resilience (Freelon et al. 2020). This visibility is essential not only for raising awareness but also for bridging geographical and social boundaries, providing a sense of belonging and solidarity to those who might otherwise feel isolated (Goswami 2018).

Furthermore, the rapid mobilization for activism that social media enables has been instrumental in advocating for societal change. Platforms like X and Facebook facilitate the rapid dissemination of information and coordination of events, exerting significant pressure on institutions to enact changes (Jackson et al. 2018). This dynamic is enhanced by social media's role as an educational resource, offering access to information about racial issues and engaging with different cultural perspectives, thereby combating stereotypes and enriching the public's understanding of complex social dynamics (Noble 2018).

Moreover, social media encourages dialogue across various demographic and ideological boundaries. Despite the inherent risks of increased polarization, the direct interactions these platforms facilitate provide opportunities for greater mutual understanding and can lead to

the reevaluation of entrenched views (Valenzuela and Reny 2020). The transparency and accountability that social media promotes are crucial in documenting incidents of racism or injustice, ensuring that these events gain the necessary visibility to prompt action and hold perpetrators accountable (Tynes et al. 2020).

Through these mechanisms, social media both mirrors and influences societal views, playing a dual role in reflecting existing prejudices and driving movements toward racial equality. These platforms are thus significant forces in the ongoing discourse on race, capable of both supporting transformative change and perpetuating divisions.

The Bad: Perpetuating polarization and outrage ouroboros

Alongside the benefits of social media are its challenges: it often solidifies divides due to algorithms that selectively amplify emotionally charged content (Anti-Defamation League 2023). This algorithmic bias towards content that provokes strong reactions creates fertile ground for misinformation, contributing to heightened societal divisions (Bail et al. 2018; Guess et al. 2018). This effect is not merely a byproduct but a central feature of how social media platforms are engineered to maximize user engagement, often at the cost of nuanced discourse.

Moreover, the amplification of moral outrage on these platforms plays a significant role in shaping the dynamics of polarization, particularly within racial politics. Digital platforms expedite and magnify the spread of outrage, transforming it into a powerful but volatile force within public discourse (Crockett 2017). The immediacy and anonymity provided by digital interactions exacerbate tensions and reinforce prejudices, pushing the boundaries of political rhetoric into more extreme territories (Tucker et al. 2018).

Outrage culture on social media often results in what has been termed 'call-out' culture, where individuals or groups are publicly shamed or attacked, typically for perceived or actual infractions of social norms or moral values. This phenomenon can lead to a hyper-vigilant environment where the fear of misstepping becomes pervasive (Ronson 2015). While call-out culture can be seen as a form of social enforcement of community standards, it can also result in over-correction and targeting individuals disproportionately without due process.

The interaction between social media algorithms and human behavior has created a feedback loop that not only reflects but also amplifies societal divisions. Studies have shown that exposure to opposing viewpoints on social media, rather than fostering understanding, often entrenches beliefs and attitudes, leading to increased polarization (Suhay et al. 2018). This phenomenon is compounded by the structure of social media networks, where users often engage in echo chambers that reinforce their own views and denigrate those of the opposition (Pariser 2011). However, this claim has critics who argue that the diversity of connections on social networks can expose individuals to a broader range of opinions than those typically encountered in offline environments (Barberá et al. 2015). According to Barberá et al. (2015), while some individuals do experience content that aligns closely with their ideological preferences, social media platforms inherently possess diverse user bases which can lead to increased exposure to conflicting viewpoints, countering the echo chamber effect.

Furthermore, the business models of these platforms, which prioritize engagement over accuracy, have facilitated the rapid spread of misinformation and conspiracy theories (Anti-Defamation League 2023). This has profound implications for political communication, particularly in contexts that are already racially charged (Allcott and Gentzkow 2017). The role of misinformation in shaping electoral outcomes and public opinion has become a significant concern, as seen in various global political events (Vosoughi et al. 2018).

In addition to facilitating misinformation, the global reach and lack of effective regulatory frameworks for social media pose challenges for combating hate speech and other forms of harmful content. Despite efforts by platforms to police content, the sheer volume and speed of information flow make it nearly impossible to control completely (Gillespie 2018). These regulatory challenges are further complicated by the international nature of the internet, where different cultural norms and legal standards apply (UNESCO 2022).

Thus, while social media can act as a catalyst for positive change and community building, its propensity to amplify the worst aspects of human discourse cannot be overlooked. Understanding and mitigating the negative impacts of social media on societal discourse, particularly in the context of racial and political polarization, remains a crucial area for ongoing research and policy development (Tufekci 2015).

The Ugly: Exacerbating enmity and engineering exclusion

The challenge of hate speech on social media platforms presents a stark manifestation of the ugliest aspects of online interactions. Hate speech not only underscores significant disparities across demographic lines but also reveals how identity and political affiliation shape the reception and impact of such toxic discourse (Solovev and Pröllochs 2022; Farrand 2023). This section explores the pervasive nature of online hate speech, the mechanisms behind its spread, and the efforts (or lack thereof) of platform governance efforts.

Hate speech in digital spaces often thrives under the cloak of anonymity, which lowers social inhibitions and encourages individuals to express opinions they might otherwise keep private (Citron 2014). This anonymity can exacerbate the intensity of hate speech, making social media a breeding ground for hostility and misinformation. Engaging with or even being exposed to hate speech can have profoundly negative consequences for the targets of such

speech and can also normalize aggressive behaviors in broader online and offline communities (Gagliardone et al. 2015).

The normalization of hate speech contributes to a recalibration of societal norms, where counter-normative expressions of intolerance can become more accepted. The effect of such normalization is not just theoretical but observable in shifts in public discourse and an increase in polarization (Mutz 2018). Moreover, interaction with uncivil discourse online does not just harm those directly involved; it has ripple effects throughout society, influencing bystanders and shaping the social fabric (Coe et al. 2014).

One of the most insidious effects of hate speech is its ability to stifle marginalized voices. Individuals from marginalized communities often face disproportionate amounts of hate speech, leading to self-censorship and withdrawal from public discourse. This exclusionary effect perpetuates existing inequalities by silencing voices crucial for diverse and inclusive dialogues. When marginalized voices are silenced, societal discourse becomes skewed, reinforcing dominant narratives and entrenching systemic biases.

Platforms like Facebook and Instagram have implemented various strategies to combat the spread of hate speech, but these efforts have met with mixed success (Farrand 2023). Since the acquisition by Elon Musk, X has seen changes that affect how hate speech is monitored, including a new paywall that limits access to data, complicating efforts by researchers to study and track changes in online behavior (Davidson et al. 2023). Meanwhile, Facebook has employed sophisticated algorithms and human moderation teams to detect and remove hate speech, though the sheer volume and subtlety of such speech pose continuous challenges (Gorwa 2020).

The effectiveness of these regulatory efforts is further complicated by the global reach of these platforms, which must navigate a maze of international laws and cultural norms (UNESCO

2022). While some regions demand stringent controls on hate speech, others advocate for maximal freedom of expression, creating a regulatory patchwork that is difficult to manage (Global Witness 2022; Amnesty International 2023).

Examining the historical and digital evolution of racial discourse reveals profound changes in political communication and societal engagement with race. These shifts reflect and influence how racial attitudes are discussed, understood, and acted upon in the public domain. Modern digital platforms, especially social media, further complicate and amplify these dynamics. Our theory aims to unravel how social media mirrors historical patterns and actively participates in creating and propagating new racial norms, fundamentally transforming the discourse.

3.4 Theory

This section proposes a theory of how social media uniquely influences the proliferation of speech that is counter-normative to the established conventions of racial equality. Social media platforms, which have been pivotal in propelling movements like Black Lives Matter and Me Too, also paradoxically facilitate the spread and organization of counter-movements. These platforms allow for what can be described as 'ideological clustering,' where individuals with racist ideologies can find and reinforce each other's beliefs, thereby forming a potent echo chamber (Bliuc et al. 2018).

Social media's role in shaping racial discourse extends beyond mere communication; it actively constructs and reconstructs the social norms pertaining to race. As Tankard and Paluck (2016) noted, the reference groups to which individuals belong significantly influence the norms they consider important. In the context of racial identity, as White Americans perceive their political influence waning, their racial identity often becomes more pronounced, leading to increased group cohesion. This cohesion, coupled with perceived external threats, may

increase the likelihood of expressing animus towards significant out-group populations, such as Black Americans.

Moreover, the interaction between individuals' perception of social norms and their personal behavior is crucial. Frequent exposure to counter-normative behavior on social media can normalize such behavior, making it more acceptable to express similar sentiments (Tankard and Paluck 2016). Social media not only allows individuals to encounter these counter-normative views but also provides a platform for their propagation among like-minded audiences, thus serving as a powerful tool for the distribution of ideologies that challenge the established norms of racial equality (Borg et al. 2021).

In synthesizing these elements, the theory posits that social media has reconfigured the landscape of racial discourse by enabling the rapid spread and reinforcement of counter-normative racial ideologies. This dynamic, driven by the unique capabilities of social media for rapid dissemination and community formation, has significant implications for how racial attitudes are formed, maintained, and changed in contemporary society. It underscores the need for a nuanced understanding of the digital platforms' role in shaping not just public discourse but also the very norms that govern societal behavior towards race.

3.4.1 Hypothesis

Building on the theoretical foundation discussed, this section outlines a specific hypothesis that seeks to empirically test the dynamics of racial discourse as influenced by social media. The theory suggests that social media serves as a powerful medium for the dissemination and amplification of racial ideologies, both normative and counter-normative. This dynamic plays a critical role in shaping societal attitudes and behaviors, particularly in how individuals perceive and interact with different racial groups.

Given the ability of social media to facilitate rapid spread and acceptance of diverse ideas, it is hypothesized that exposure to counter-normative racial comments on these platforms will significantly impact public perception and attitudes. These interactions on social media do not merely reflect existing social sentiments but actively shape them, often reinforcing or challenging the prevailing norms of racial equality. The specific hypothesis proposed is:

> *Hypothesis 1: Engagement with counter-normative racial comments on social media platforms leads to an increase in negative racial attitudes among users and more favorable views of the announcer, compared to tweets without counter-normative comments.*

This hypothesis stems from the observation that social media often blurs the lines between private opinion and public endorsement, allowing users to engage with content that they might not encounter or endorse in other settings. The ideological clustering on social media platforms can create echo chambers that reinforce specific worldviews, including those that are racially charged or counter-normative. Exposure to such content is likely to influence users' attitudes, potentially normalizing views that are divergent from mainstream societal norms.

Furthermore, this hypothesis aligns with the notion that social media can act as a 'megaphone' for minority opinions, providing a platform where counter-normative views can gain visibility and traction. This visibility can affect how individuals perceive the acceptability of certain attitudes and behaviors, shifting public opinion and possibly leading to greater polarization.

Testing this hypothesis will involve examining the relationship between exposure to racial discourse on social media and subsequent changes in racial attitudes. This approach will help illuminate the complex interplay between digital media consumption and the evolution of

racial norms in contemporary society, offering insights into how digital platforms could be better managed to foster more inclusive and respectful public discourse.

3.5 Research design

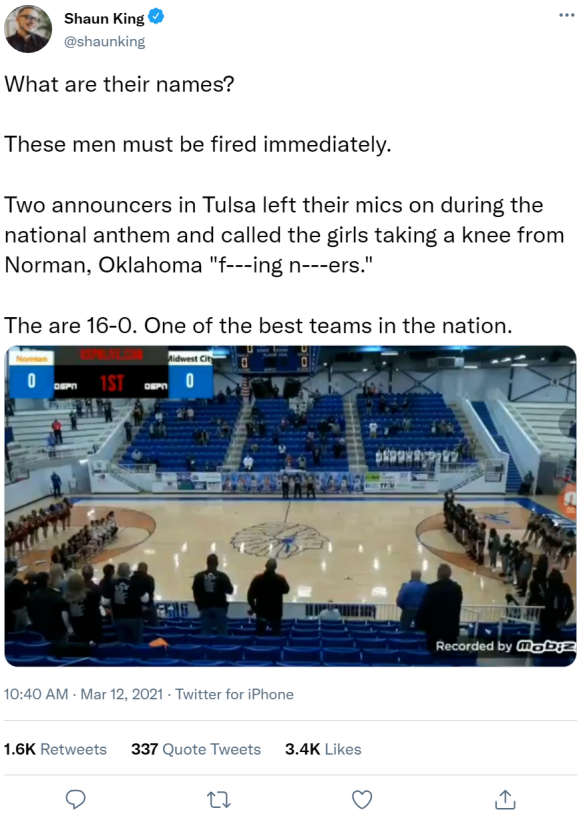
The primary stimulus for this experiment is a tweet, posted on Twitter on March 12, 2021, which involves a sports announcer using a racial slur during a high school basketball game. This incident, depicted in Figure 3.2, serves as a relevant example of speech that breaches norms of racial equality and provokes varied public reactions.

Participants were assigned to one of five treatment groups, each designed to explore different aspects of interaction with a racially charged tweet. The groups were carefully structured to assess how various contexts of exposure to the tweet influence individual reactions and attitudes.

In this study, the control group functioned as a foundational baseline, where participants were not exposed to the tweet or its replies, establishing a clear comparative framework for analyzing the influence of direct exposure to the tweet content. Participants in the 'Tweet-only group' viewed the tweet absent any replies, which isolates and clarifies the tweet's inherent impact without the moderating effects of community interaction.

The experiment further dissected the role of community feedback through two distinct groups: the 'Normative replies group' and the 'Counter-normative replies group'. The former encountered the tweet alongside replies that condemned the racial slur, probing the effectiveness of normative, corrective feedback on altering viewer perceptions. In contrast, the latter group interacted with replies that endorsed the use of the slur, exploring how supportive counter-normative comments influence attitudes towards such divisive content.

Figure 3.2: Treatment tweet

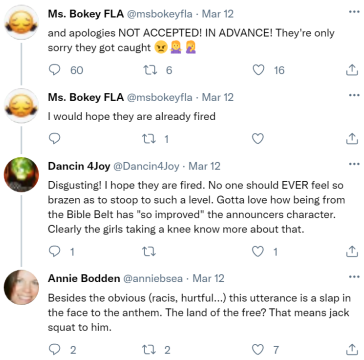


Illustrates a tweet by Shaun King, highlighting an incident involving racial comments made by sports announcers. The tweet calls for the identification and immediate termination of the announcers who were caught making derogatory remarks about a girls' basketball team from Norman, Oklahoma, during the national anthem. Demonstrates the Tweet-only Condition: Participants in this condition were shown only the original tweet without any responses, to assess their reactions based solely on the tweet itself.

Additionally, the 'Mixed replies group' provided participants with a more representative social media experience, exposing them to a blend of both supportive and condemning replies. This setup mimics the multifaceted nature of actual social media interactions, where users frequently confront a spectrum of opinions that may simultaneously guide and complicate their perceptions and attitudes.

Figure 3.3: Treatment conditions

(a) Mixed replies



(b) Normative replies

(c) Counter-normative replies

Displayed are the three distinct treatment conditions used to examine how different types of community feedback influence individual perceptions and reactions to racially charged content: (a) Mixed replies: This condition includes a combination of normative and counter-normative responses to the initial tweet, offering a realistic scenario of mixed social feedback. (b) Normative replies: This group consists solely of responses that condemn the use of a racial slur, highlighting the community's enforcement of anti-discriminatory norms. (c) Counter-normative replies: This setting includes replies that support or endorse the racial slur, providing a perspective where the normative stance against discrimination is challenged.

This array of treatment groups serves as the primary independent variable of the study, supplemented by an interaction with pre-treatment measures of racial resentment to comprehensively assess the effects on participant responses.

Opinions were operationalized through participant responses to six statements concerning the appropriateness of the announcer's language and its potential consequences. These responses were gathered using a five-point Likert scale that addressed issues from free speech rights

to the acceptability of racial slurs, with higher values indicating more counter-normative opinions. A composite opinion score was then derived by averaging these responses.

Racial attitudes were evaluated using various scales. The F.I.R.E. battery, assessing fear, acknowledgment of institutionalized racism, and empathy, required participants to rate their agreement with four statements on a four-point scale, where higher scores denoted more negative racial attitudes (DeSante and Smith 2020). The racial resentment scale, a prevalent tool in political science, measures anti-Black affect, beliefs about work ethic, and denial of discrimination. The exact questions for the F.I.R.E battery and racial resentment scale can be seen in the survey instrument on page 166 of Appendix C.

Explicit racism was gauged using the dehumanization scale and the social distance measure, both post-treatment, with lower scores reflecting greater dehumanization and increased social distance, respectively.

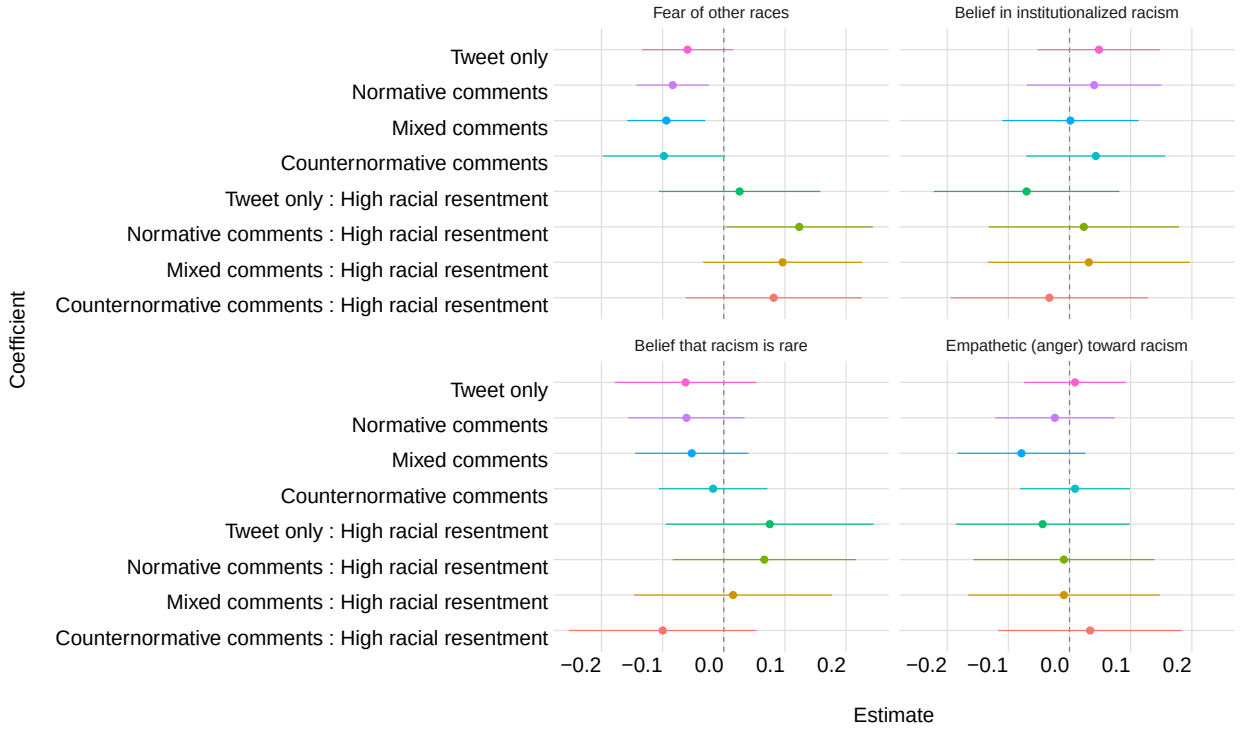
3.6 Study

Chapter 3 explores the impact of exposure to various types of social media comments on racial attitudes, particularly how these exposures interact with pre-existing levels of racial resentment. This analysis provides insights into the dynamics of online interactions and their potential effects on societal views. More detailed results including additional controls and models without interaction terms are available on page 157.

3.6.1 Results

The first half of this analysis focuses on the F.I.R.E. battery, which assesses participants' **fear** of other races, recognition of **institutional** racism and the extent to which they believe such **racism** exists, and **empathy** towards racial issues.

Figure 3.4: Impact of post/comments and racial resentment level on F.I.R.E battery among participants



Illustrates the standardized regression coefficients for four different models examining the F.I.R.E battery. Each panel represents a different dependent variable: "Fear of other races", "Belief in institutionalized racism", "Belief that racism is rare", and "Epathetic (anger) toward racism". Higher values indicate more counter-normative opinions about race, so higher values of belief in institutionalized racism and empathy toward racism indicate less belief in racism and less empathy toward racism. For clarity, control variables such as political identification (PID), Republican PID, bachelor's degree, age categories (30-44 years, 46-49 years, and 60+ years), male, and income were included in the regression analysis but are not shown in the plot. See page 158 of Appendix C for regressions.

The results depicted in Figure 3.4 demonstrate the nuanced effects of social media content and personal racial resentment levels on various racial attitudes among participants, using the F.I.R.E battery as a framework. This battery assesses attitudes across several dimensions: Fear of other races, Belief in institutionalized racism, Belief that racism is rare, and Empathetic reactions toward racism.

For the dimension assessing fear of other races, the data suggests minimal impact from the types of social media comments—whether tweet-only, normative, mixed, or counter-normative. The coefficients are all relatively close to zero, implying that exposure to different types of social media discourse about race does not significantly influence participants' fear levels towards other races. This outcome indicates that the expression of fear or security in racial contexts may be more deeply rooted and less susceptible to change through brief social media interactions.

Regarding belief in institutionalized racism, the results are similarly subtle, with only tweet-only comments showing a slight negative influence. This could suggest that direct exposure to racial content without additional commentary slightly decreases belief in widespread systemic racism. However, other comment types do not show significant effects, pointing to the complexity of how individuals process information related to institutional racism and how deeply held these beliefs are.

The dimension of belief that racism is rare shows a similarly muted pattern, with none of the coefficients indicating significant shifts in belief due to different types of comments. This could reflect an entrenched skepticism or acknowledgment of racism's prevalence, which is not easily swayed by single instances of exposure to social media content.

Lastly, the aspect of empathetic reactions toward racism shows no significant changes based on the type of comments participants were exposed to. This might suggest that empathy, as

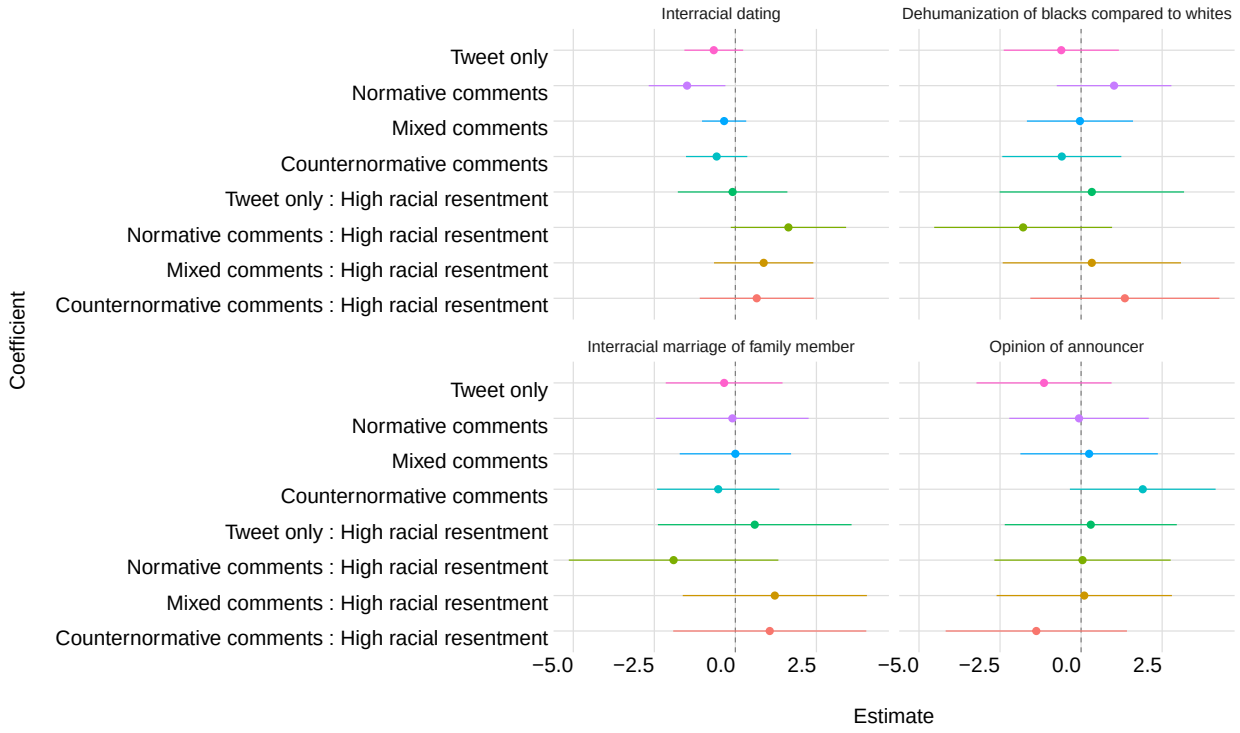
an emotional response to racism, is deeply ingrained and not easily altered by the nature of social media discourse alone.

Together, these results underscore a broader implication: while social media undoubtedly plays a role in shaping public discourse on race, its ability to alter deeply held racial attitudes through brief exposures to content might be limited. It suggests that racial attitudes, whether fear, belief in institutional racism, skepticism about its rarity, or empathy towards those affected, are likely shaped by more prolonged and complex social interactions and not merely by transient social media engagements. This finding points to the resilience of existing racial attitudes and the challenges faced in efforts to shift these perceptions through online platforms.

Moving to Figure 3.5, the study also examined broader racial attitudes and opinions through several other models. In the analysis, we observe the impacts of social media interactions and individual racial resentment on shaping racial attitudes. The standardized regression coefficients explore four primary outcomes: attitudes toward interracial dating, perceptions of dehumanization, comfort with interracial marriage, and opinions of the announcer. This detailed breakdown helps in understanding how social media discourse interacts with personal biases to influence societal views on race.

For attitudes toward interracial dating and marriage, as illustrated in Figure 3.5, the coefficients for interactions between comment types and high racial resentment—across both dating and marriage—are generally small and not statistically significant. This suggests that exposure to different types of social media comments, whether normative, mixed, or counter-normative, does not substantially alter these attitudes, even among individuals with high racial resentment. This finding could imply that such views are deeply entrenched and not easily swayed by social media interactions alone.

Figure 3.5: Impact of post/comments and racial resentment level on attitudes and opinions of participants



Illustrates the standardized regression coefficients for four different models examining attitudes towards race. Each panel represents a different dependent variable: "Interracial dating", "Dehumanization of blacks compared to whites", "Interracial marriage of a family member", and "Opinion of the announcer". Higher values indicate more counter-normative opinions about race, such as support for the announcer. For clarity, control variables such as political identification (PID), Republican PID, bachelor's degree, age categories (30-44 years, 46-49 years, and 60+ years), male, and income were included in the regression analysis but are not shown in the plot. The findings were standardized to allow comparison across different models using the formula: (estimate - mean(estimate)) / sd(estimate). See page 162 of Appendix C for regressions.

Regarding dehumanization, the data reveals a similar pattern. None of the social media comment types show significant impacts on how participants perceive the dehumanization of blacks compared to whites. This lack of significant change, including in the context of high racial resentment, underscores the resilience of established racial attitudes and the complex challenges in addressing racial biases through digital platforms.

When examining opinions about the announcer, a slightly more nuanced picture emerges from the regression model. While still generally small, the coefficients indicate that negative views about the announcer do not significantly change based on the type of comments viewed. However, the presence of high racial resentment slightly modifies this pattern, although not enough to show a statistically significant effect.

Overall, the data summarized in Figure 3.5 suggests that while social media has the potential to expose individuals to diverse viewpoints, its impact on deeply held racial attitudes and perceptions is limited. This resilience against change points to the entrenched nature of racial attitudes and highlights the need for more comprehensive approaches to influence public opinion on race, beyond mere exposure to varied social media content.

3.6.2 Discussion

This study explored the impact of social media comments on racial attitudes by exposing participants to different types of social media interactions. The results, largely indicating null effects across various measures of racial attitudes, suggest that brief exposures to social media content do not significantly alter deep-seated racial beliefs. This finding points to the resilience of racial attitudes, which appear to be shaped by more enduring and complex social influences than a single post or series of comments.

It may be the case that the enduring nature of racial attitudes might require more profound or repeated exposures to alternative viewpoints to initiate change. Social media interactions, despite their intensity and immediacy, might not provide sufficient depth or continuity to influence these deeply ingrained beliefs effectively. Alternatively, the format of the experiment—focusing on brief interactions—might not replicate the ongoing, cumulative exposure to social media that shapes public opinions over time.

The study's design also presents limitations that could affect the interpretation of the results. For example, the 'mixed replies' condition included only two comments, potentially limiting the representation of a genuine social media environment, where a broader array of responses might be encountered. This could affect the ecological validity of the findings, as the controlled nature of the experiment does not fully capture the complexity and variability of real-world social media interactions.

Furthermore, the survey's focus on X interactions might not generalize across all social media platforms, which vary in user demographics, content presentation, and interaction styles. This specificity could limit the broader applicability of the findings to other contexts where racial attitudes are discussed and formed.

Additional limitations include potential biases in participant selection and response accuracy. Participants who choose to engage in a study about racial attitudes might not represent the broader population's views, and self-reported measures can be subject to social desirability biases, particularly concerning sensitive topics like race.

3.7 Conclusion

The findings of this study, situated within the broader literature on social media and racial discourse, highlight the complexities of influencing racial attitudes through digital platforms. While current research often emphasizes the negative impacts of social media, this study reveals the resilience of normative racial attitudes.

These findings are significant because they suggest that negative encounters on social media do not have immediate and profound impacts on deeply held beliefs. This resilience offers a

level of stability in societal attitudes, which can be seen as a positive outcome in the context of avoiding rapid, negative shifts in public sentiment.

Beyond academia, these insights are valuable for policymakers, social media developers, and community leaders. They suggest that while social media might not quickly change deeply held beliefs, it also does not easily amplify negative changes in attitudes. This stability can be leveraged to design more effective, long-term strategies for addressing social issues. Understanding the capabilities and limits of these tools in shaping public discourse is crucial for fostering a more inclusive and understanding society.

Afterword

As this journey through the intricate landscape of cognitive biases, misinformation, and racial discourse in the digital age concludes, a blend of insights and new questions comes to light. The exploration began with an examination of argument congruency bias, revealing the intricate dance between logic and pre-existing beliefs in how individuals evaluate information. It was discovered that while people can distinguish between strong and weak arguments, they are inevitably influenced by their own convictions, painting a picture of human reasoning that is both capable and flawed.

Moving deeper into the realm of misinformation, the investigation ventured into the world of culturally-relevant corrections on social media. Here, a surprising twist emerged. Contrary to expectations, the effectiveness of corrections was not necessarily enhanced by cultural congruence. Instead, the perceived expertise and relevance of the source emerged as crucial factors, challenging assumptions about in-group bias and highlighting the nuanced nature of trust and credibility in information processing.

The final chapter extended this exploration of social media's influence, focusing specifically on its impact on racial attitudes. As the effects of varied social media content on deeply held racial beliefs were examined, an unexpected stillness was encountered. The brief exposures

to different types of comments, which were hypothesized to sway racial attitudes, instead revealed the steadfast nature of beliefs.

Throughout this narrative, a common thread emerges: the complexity of human information processing. From argument evaluation to misinformation correction to racial attitude formation, a recurring theme of cognitive resilience is evident. The mind, it seems, is not easily swayed by single exposures or simple interventions. This resilience serves as both a shield against misinformation and counter-normative speech and a barrier to the correction of mistaken beliefs.

Bibliography

Aïmeur, Esma , Sabrine Amri, and Gilles Brassard (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining 13*(30). Accessed: 87k; Altmetric: 163; Mentions: 31.

Allcott, Hunt and Matthew Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*(2), 211–236.

Allport, Gordon W. (1954). *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Amazeen, Michelle A (2015). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly 92*(1), 27–48.

Amazeen, Michelle A (2020). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly 97*(1), 204–222.

Amnesty International (2023). Contemporary forms of hate speech online. https://www.amnesty.org/ar/wp-content/uploads/2023/07/IOR4069862023ENGLISH.pdf.

Anti-Defamation League (2023). Two studies: Social media algorithms fuel online hate. https://www.adl.org/resources/report/bad-worse-amplification-and-auto-generation-hate.

Bail, Christopher A , Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Banas, John A and Stephen A Rains (2010). A meta-analysis of the efficacy of inoculation theory. *Communication Monographs 77*(3), 281–311.

Banks, Antoine J (2014). *Anger and racial politics: The emotional foundation of racial attitudes in America*. Cambridge University Press.

Banks, Antoine J and Nicholas A Valentino (2012). Emotional substrates of white racial attitudes. *American Journal of Political Science 56*(2), 286–297.

Barberá, Pablo , John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science 26*(10), 1531–1542.

Bargh, John A , Peter M Gollwitzer, Annette Lee-Chai, Kimberly Barndollar, and Roman Trötschel (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology 81*(6), 1014.

Benkler, Yochai , Casey Tilton, Bruce Etling, Hal Roberts, Justin Clark, Robert Faris, Jonas Kaiser, and Carolyn Schmitt (2020). Mail-in voter fraud: Anatomy of a disinformation campaign. *Berkman Center Research Publication* (2020-6).

Berinsky, Adam J , Gregory A Huber, and Gabriel S Lenz (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political analysis 20*(3), 351–368.

Berinsky, Adam J , Michele F Margolis, and Michael W Sances (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science 47*(2), 241–262.

Binder, Jens , Hanna Zagefka, Rupert Brown, Friedrich Funke, Thomas Kessler, Amélie Mummendey, Andreas Maquil, Stephanie Demoulin, and Jacques-Philippe Leyens (2009). Does contact reduce prejudice or does prejudice reduce contact? a longitudinal test of the contact hypothesis amongst majority and minority groups in three european countries. *Journal of Personality and Social Psychology 96*(4), 843–856.

Bisgaard, Martin (2015). Bias will find a way: Economic perceptions, attributions of blame, and partisan-motivated reasoning during crisis. *The Journal of Politics 77*(3), 849–860.

Bizer, George Y , Jon A Krosnick, Richard E Petty, Derek D Rucker, and S Christian Wheeler (2000). Need for cognition and need to evaluate in the 1998 national election survey pilot study. *National election studies report*.

Blanchard, Fay A. , Christian S. Crandall, John C. Brigham, and Leigh Ann Vaughn (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology 79*(6), 993–997.

Bliuc, Ana-Maria , Nicholas Faulkner, Andrew Jakubowicz, and Craig McGarty (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior 87*, 75–86.

Bobo, Lawrence and James R Kluegel (1998). Race, interests, and beliefs about affirmative action: Unanswered questions and new directions. *American Behavioral Scientist 41*(7), 985–1003.

Bolsen, Toby , James N Druckman, and Fay Lomax Cook (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior 36*(2), 235–262.

Borg, Kim , Jo Lindsay, and Jim Curtis (2021). When news media and social media meet: How facebook users reacted to news stories about a supermarket plastic bag ban. *New Media & Society 23*(12), 3574–3592.

Brewer, Marilynn B (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin 86*(2), 307.

Brewer, Marilynn B (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues 55*(3), 429–444.

Cheung, Christy M.K. , Randy Yee Man Wong, and Tommy K.H. Chan (2021). Online disinhibition: conceptualization, measurement, and implications for online deviant behavior. *Industrial Management & Data Systems 121*(1), 48–64.

Chong, Dennis and James N Druckman (2007). Framing theory. *Annual Review of Political Science 10*, 103–126.

Citron, Danielle Keats (2014). *Hate Crimes in Cyberspace*. Harvard University Press.

CNN Wire Staff (2011). Email depicts obama as chimpanzee. https://www.cnn.com/2011/POLITICS/04/19/obama.chimpanzee.email/index.html.

Cobb, Jelani (2016). *The Matter of Black Lives*. The New Yorker.

Coe, Kevin , Kate Kenski, and Stephen A. Rains (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication 64*(4), 658–679.

Compton, Josh (2013). Inoculation theory. *The Sage handbook of persuasion: Developments in theory and practice 2*, 220–236.

Compton, Josh , Ben Jackson, and James A Dimmock (2016). Persuading others to avoid persuasion: Inoculation theory and resistant health attitudes. *Frontiers in psychology 7*, 122.

Cook, John , Naomi Oreskes, Peter T Doran, William RL Anderegg, Bart Verheggen, Ed W Maibach, J Stuart Carlton, Stephan Lewandowsky, Andrew G Skuce, Sarah A Green, et al. (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters 11*(4).

Coppock, Alexander (2019). Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods 7*(3), 613–628.

Corr, Méabh , Jaimie McMullen, Philip J. Morgan, Alyce Barnes, and Elaine M. Murtagh (2019, Nov). Supporting our lifelong engagement: Mothers and teens exercising (sole mates); a feasibility trial. *Women & Health*, 1–18.

Crockett, Molly J. (2017). Moral outrage in the digital age. *Nature Human Behaviour 1*(11), 769–771.

Custers, Ruud and Henk Aarts (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science 329*(5987), 47–50.

Davidson, Brittany I. , Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen, and Alicia G. Cork (2023). Platform-controlled social media apis threaten open science. *Nature Human Behaviour 7*, 2054–2057.

Davis, Jason L and Kevin R Binning (2017). Building a more accurate political belief system: The effects of epistemic motivation on the integration of political information. *Social Psychological and Personality Science 8*(8), 852–862.

DeSante, Christopher D and Candis Watts Smith (2020). Fear, institutionalized racism, and empathy: The underlying dimensions of whites' racial attitudes. *PS: Political Science & Politics 53*(4), 639–645.

Devine, Patricia G (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology 56*(1), 5.

Doosje, Bertjan , Russell Spears, and Naomi Ellemers (2002). Social identity as both cause and effect: The development of group identification in response to anticipated and actual changes in the intergroup status hierarchy. *British Journal of Social Psychology 41*(1), 57–76.

Dovidio, John F and Samuel L Gaertner (2004). Aversive racism. *Advances in experimental social psychology 36*, 4–56.

Druckman, James N (2010). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication 60*(1), 142–156.

Ecker, Ullrich KH and Li Chang Ang (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology 40*(2), 241–260.

Ecker, Ullrich KH , Stephan Lewandowsky, Lisa EP Chang, and Rekha Pillai (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied 20*(4), 323.

Effron, Daniel A and Eric D Knowles (2015). Entitativity and intergroup bias: How belonging to a cohesive group allows people to express their prejudices. *Journal of personality and social psychology 108*(2), 234.

Farrand, Benjamin (2023). 'is this a hate speech?'the difficulty in combating radicalisation in coded communications on social media platforms. *European Journal on Criminal Policy and Research 29*(3), 477–493.

Feinberg, Matthew and Robb Willer (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin 41*(12), 1665–1681.

Flynn, DJ , Brendan Nyhan, and Jason Reifler (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology 38*, 127–150.

Freelon, Deen , Charlton D. McIlwain, and Meredith Clark (2020). Beyond the hashtags: ferguson, blacklivesmatter, and the online struggle for offline justice. *American Behavioral Scientist 64*(7), 1012–1031.

Gagliardone, Iginio , Danit Gal, Thiago Alves, and Gabriela Martinez (2015). *Countering Online Hate Speech*. UNESCO Publishing.

Gaines, Brian J , James H Kuklinski, Paul J Quirk, Buddy Peyton, and Jay Verkuilen (2007). Same facts, different interpretations: Partisan motivation and opinion on iraq. *Journal of Politics 69*(4), 957–974.

Garrett, R. Kelly , Erik C. Nisbet, and Emily K. Lynch (2013). Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory. *Journal of Communication 63*(4), 617–637.

Gilens, Martin (1996). "race coding" and white opposition to welfare. *American Political Science Review 90*(3), 593–604.

Gillespie, Tarleton (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Global Witness (2022). Facebook's role in amplifying hate speech in ethiopia. https://www.globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/.

Gollwitzer, Peter M. and John A. Bargh (2005). Automaticity in goal pursuit. In A. J. Elliot and C. S. Dweck (Eds.), *Handbook of Competence and Motivation*, pp. 624–646. Guilford Publications.

Gorwa, Robert (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7*(1), 205395172091977.

Goswami, Manash Pratim (2018). Social media and hashtag activism. *Liberty dignity and change in journalism 2017*.

Grant, Peter R and Rupert Brown (1995). From ethnocentrism to collective protest: Responses to relative deprivation and threats to social identity. *Social Psychology Quarterly*, 195–212.

Greene, Steven (2004). Social identity theory and party identification. *Social science quarterly 85*(1), 136–153.

Groenendyk, Eric and Yanna Krupnikov (2020). What motivates reasoning? a theory of goal-dependent political evaluation. *American Journal of Political Science*.

Guess, Andrew , Jonathan Nagler, and Joshua Tucker (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances 5*(1).

Guess, Andrew , Brendan Nyhan, and Jason Reifler (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 u.s. presidential campaign. *European Research Council*.

Guess, Andrew M and M Lerner (2020). Digital media and political misinformation. *Handbook of digital media and democracy*.

Guess, Andrew M , Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar (2020). A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences 117*(27), 15536–15545.

Halttunen, Karen (2008). The obama cover controversy. *OAH Newsletter 36*(3). Available at: `https://www.oah.org/insights/archive/the-obama-cover-controversy/`.

Hameleers, Michael and Toni G. L. A. van der Meer (2019, Jan). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, 009365021881967.

Hart, P Sol and Erik C Nisbet (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication research 39*(6), 701–723.

Hassin, Ran R , Henk Aarts, Baruch Eitam, Ruud Custers, and Tali Kleiman (2009). Non-conscious goal pursuit and the effortful control of behavior. *Oxford handbook of human action*, 549–566.

Hewstone, Miles , Mark Rubin, and Hazel Willis (2002). *Intergroup bias*. Annual Reviews.

Hill, Seth J (2017). Learning together slowly: Bayesian learning about political facts. *The Journal of Politics 79*(4), 1403–1418.

Hobbs, Renee (1999). Teaching media literacy: Yo! are you hip to this? *Social Education 63*(3), 123–129.

Hopkins, Daniel J , John Sides, and Jack Citrin (2019). The muted consequences of correct information about immigration. *The Journal of Politics 81*(1), 315–320.

Huddy, Leonie and Stanley Feldman (2009). On assessing the political effects of racial prejudice. *Annual Review of Political Science 12*, 423–447.

Hutchings, Vincent L , Hanes Walton Jr, and Andrea Benjamin (2010). The impact of explicit racial cues on gender differences in support for confederate symbols and partisanship. *The Journal of Politics 72*(4), 1175–1188.

Iyengar, Shanto , Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science 22*, 129–146.

Jackson, Sarah J. , Moya Bailey, and Brooke Foucault Welles (2018). *HashtagActivism: Networks of Race and Gender Justice*. MIT Press.

James, Oliver and Gregg G Van Ryzin (2016). Motivated reasoning about public performance: An experimental study of how citizens judge the affordable care act. *Journal of Public Administration Research and Theory 27*(1), 197–209.

Jardina, Ashley (2019). *White identity politics*. Cambridge University Press.

Jeong, Se-Hoon , Hyunyi Cho, and Yoori Hwang (2009). Media literacy interventions: A meta-analytic review. *Journal of Communication 59*(3), 454–472.

Jervis, Robert (2017). *Perception and misperception in international politics: New edition*. Princeton University Press.

Kahne, Joseph and Benjamin Bowyer (2017). Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal 54*(1), 3–34.

Kahneman, Daniel (2011). *Thinking, fast and slow*. Macmillan.

Kinder, Donald R and Cindy D Kam (2010). *Us against them: Ethnocentric foundations of American opinion*. University of Chicago Press.

Kuklinski, James H. and Paul J. Quirk (2000). Reconsidering the rational public: Cognition, heuristics, and mass opinion. *Elements of reason: Cognition, choice, and the bounds of rationality*, 153–182.

Kunda, Ziva (1990). The case for motivated reasoning. *Psychological bulletin 108*(3), 480.

Lau, Richard R (1989). Individual and contextual influences on group identification. *Social psychology quarterly*, 220–231.

Lazer, David MJ , Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. (2018). The science of fake news. *Science 359*(6380), 1094–1096.

Lewandowsky, Stephan , Ullrich KH Ecker, and John Cook (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition 6*(4), 353–369.

Limbaugh, Rush (2007). Barack the Magic Negro. Radio segment aired on The Rush Limbaugh Show.

Lodge, Milton and Charles S Taber (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science 50*(3), 755–769.

Lodge, Milton and Charles S Taber (2013). *The rationalizing voter*. Cambridge University Press.

Lodge, Milton and Charles S Taber (2016). The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Political Psychology 37*, 61–85.

López, Ian Haney (2014). *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. Oxford University Press.

Lopez, Ian Haney (2016). *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. Oxford University Press.

Los Angeles Times (2010). Tea party controversies: Racially charged elements at rallies. https://articles.latimes.com/2010/mar/29/nation/la-na-tea-party-qa30-2010mar30.

Margolin, Drew B , Aniko Hannak, and Ingmar Weber (2018). Political fact-checking on twitter: When do corrections have an effect? *Political Communication 35*(2), 196–219.

Marien, Hans , Ruud Custers, Ran R Hassin, and Henk Aarts (2012). Unconscious goal activation and the hijacking of the executive function. *Journal of personality and social psychology 103*(3), 399.

McGuire, William J (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in experimental social psychology 1*, 191–229.

Mendelberg, Tali (2001). *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton University Press. ISBN: 9780691070711.

Mullen, Brian , Rupert Brown, and Colleen Smith (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European journal of social psychology 22*(2), 103–122.

Mullinix, Kevin J , Thomas J Leeper, James N Druckman, and Jeremy Freese (2015). The generalizability of survey experiments. *Journal of Experimental Political Science 2*(2), 109–138.

Mutz, Diana C. (2018). Status threat, not economic hardship, explains the 2016 presidential vote. *Proceedings of the National Academy of Sciences 115*(19), E4330–E4339.

Newport, Frank (2018, Jun). Democrats racially diverse; republicans mostly white.

Nir, Lilach (2011). Motivated reasoning and public opinion perception. *Public Opinion Quarterly 75*(3), 504–532.

Noble, Safiya Umoja (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Nyhan, Brendan and Jason Reifler (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior 32*(2), 303–330.

Nyhan, Brendan and Jason Reifler (2015). Displacing misinformation about events: An experimental test of causal corrections. *Journal of Experimental Political Science 2*(1), 81–93.

Nyhan, Brendan , Jason Reifler, Sean Richey, and Gary L Freed (2013). The hazards of correcting myths about health care reform. *Medical care 51*(2), 127.

Pager, Devah (2007). Marked: Race, crime, and finding work in an era of mass incarceration. *Contemporary Sociology 36*(2), 186–187.

Paluck, Elizabeth Levy , Hana Shepherd, and Peter M Aronow (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences 113*(3), 566–571.

Pariser, Eli (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Parks, Lisa (2010). *Michelle Obama: First Lady of Hope*. Lyons Press. ISBN: 9781599215211.

Pennycook, Gordon , Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature 592*(7855), 590–595.

Pennycook, Gordon and David G Rand (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition 188*, 39–50.

Pettigrew, Thomas F. and Linda R. Tropp (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology 90*(5), 751–783.

Petty, Richard E and John T Cacioppo (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion*, pp. 1–24. Springer.

Pew Research Center (2020). Social media served as an important outlet for black americans in 2020. https://www.pewresearch.org/short-reads/2020/12/11/social-media-continue-to-be-important-political-outlets-for-black-americans/.

Pfau, Michael , Bobi Ivanov, Brian Houston, and Michel Haigh (2007). Inoculation theory. *The Sage Handbook of Persuasion: Developments in Theory and Practice*, 133–146.

Piston, Spencer (2010). How explicit racial prejudice hurt obama in the 2008 election. *Political Behavior 32*(4), 431–451.

Plant, E. Ashby and Patricia G. Devine (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology 75*(3), 811–832.

Poston, Dudley and Rogelio Sáenz (2020, Jan). The us white majority will soon disappear forever.

Prior, Markus , Gaurav Sood, Kabir Khanna, et al. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science 10*(4), 489–518.

Quinley, Harold E (2009). *Anti-Americanism: Critiques at Home and Abroad, 1965–1990*. Greenwood Publishing Group.

Redlawsk, David P. (2002). Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *Journal of Politics 64*(4), 1021–1044.

Redlawsk, David P , Andrew JW Civettini, and Karen M Emmerson (2010). The affective tipping point: Do motivated reasoners ever "get it"? *Political Psychology 31*(4), 563–593.

Reid, Fraser (1987). *Rediscovering the social group: A self-categorization theory*. Wiley Online Library.

Reny, Tyler T , Ali A Valenzuela, and Loren Collingwood (2020). "no, you're playing the race card": Testing the effects of anti-black, anti-latino, and anti-immigrant appeals in the post-obama era. *Political Psychology 41*(2), 283–302.

Ronson, Jon (2015). *So You've Been Publicly Shamed*. New York: Riverhead Books.

Russett, Bruce (1993). Grasping the democratic peace: Principles for a post-cold war world. *Princeton University Press*.

Scheufele, Dietram A (2014). Science communication as political communication. *Proceedings of the National Academy of Sciences 111*(Supplement 4), 13585–13592.

Schuck, Andreas RT et al. (2017). Media malaise and political cynicism. *The international encyclopedia of media effects*, 1–19.

Sears, David O (1988). Symbolic racism. In *Eliminating racism*, pp. 53–84. Springer.

Serwer, Adam (2019, August). The white nationalists are winning. https://www.theatlantic.com/ideas/archive/2018/08/the-battle-that-erupted-in-charlottesville-is-far-from-over/567167/?curator=MediaREDEF.

Sherif, Muzafer , Carolyn W Sherif, et al. (1961). Intergroup conflict and cooperation: The robbers cave experiment. *Norman, OK: University Book Exchange*.

Sidanius, Jim and Felicia Pratto (1996). Social dominance theory: A new synthesis. *Political Psychology 17*(4), 758–760.

Slothuus, Rune and Claes H De Vreese (2010). Political parties, motivated reasoning, and issue framing effects. *The Journal of Politics 72*(3), 630–645.

Smeesters, Dirk , S Christian Wheeler, and Aaron C Kay (2010). Indirect prime-to-behavior effects: The role of perceptions of the self, others, and situations in connecting primed constructs to social behavior. In *Advances in experimental social psychology*, Volume 42, pp. 259–317. Elsevier.

Sniderman, Paul M , Thomas Piazza, Philip E Tetlock, and Ann Kendrick (1991). The new racism. *American Journal of Political Science*, 423–447.

Solovev, Kirill and Nicolas Pröllochs (2022). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *arXiv*.

Stanovich, Keith E and Richard F West (2000). *Individual differences in reasoning: Implications for the rationality debate?* Psychology Press.

Stephan, Walter G , Kurt A Boniecki, Oscar Ybarra, Ann Bettencourt, Kelly S Ervin, Linda A Jackson, Penny S McNatt, and C Lausanne Renfro (2002). The role of threats in the racial attitudes of blacks and whites. *Personality and Social Psychology Bulletin 28*(9), 1242–1254.

Strack, Fritz and Roland Deutsch (2011). A theory of impulse and reflection. *Handbook of theories of social psychology*, 97–117.

Suhay, Elizabeth , Emily Bello-Pardo, and Brianna Maurer (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics 23*(1), 95–115.

Sunstein, Cass R (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Swire, Briony , Adam J Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker (2017). Processing political misinformation: Comprehending the trump phenomenon. *Royal Society open science 4*(3), 160802.

Taber, Charles S , Damon Cann, and Simona Kucsova (2009). The motivated processing of political arguments. *Political Behavior 31*(2), 137–155.

Taber, Charles S. and Milton Lodge (2006a). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science 50*(3), 755–769.

Taber, Charles S and Milton Lodge (2006b). Motivated skepticism in the evaluation of political beliefs. *American journal of political science 50*(3), 755–769.

Tajfel, Henri (1974). Social identity and intergroup behaviour. *Information (International Social Science Council) 13*(2), 65–93.

Tajfel, Henri , John C Turner, William G Austin, and Stephen Worchel (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader 56*, 65.

Tandoc Jr, Edson C. , Zheng Wei Lim, and Richard Ling (2018). Defining 'fake news': A typology of scholarly definitions. *Digital Journalism 6*(2), 137–153.

Tankard, Margaret E and Elizabeth Levy Paluck (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review 10*(1), 181–211.

Tesler, Michael (2012). The spillover of racialization into health care: How president obama polarized public opinion by racial attitudes and race. *American Journal of Political Science 56*(3), 690–704.

Tesler, Michael (2018). Priming predispositions and changing policy positions: An account of when mass opinion is primed or changed. *American Journal of Political Science 62*(4), 806–824.

The Brennan Center (2022, January). The impact of voter suppression on communities of color. `https://www.brennancenter.org/our-work/research-reports/impact-voter-suppression-communities-color`. Accessed: 2023-04-09.

Thorson, Emily (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication 33*(3), 460–480.

Tucker, Joshua A , Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Tufekci, Zeynep (2015). Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colorado Technology Law Journal 13*(2), 203–217.

Tynes, Brendesha M , Devin English, Juan Del Toro, Naila A Smith, Fantasy T Lozada, and David R Williams (2020). Trajectories of online racial discrimination and psychological functioning among african american and latino adolescents. *Child development 91*(5), 1577–1593.

UNESCO (2022). Online hate speech and disinformation: Regulatory challenges. https://unesdoc.unesco.org/ark:/48223/pf0000387339.

Valentino, Nicholas A , Vincent L Hutchings, and Ismail K White (2002). Cues that matter: How political ads prime racial attitudes during campaigns. *American Political Science Review*, 75–90.

Valentino, Nicholas A , Fabian G Neuner, and L Matthew Vandenbroek (2018). The changing norms of racial political rhetoric and the end of racial priming. *The Journal of Politics 80*(3), 757–771.

Valenzuela, Ali A and Tyler T Reny (2020). Evolution of experiments on racial priming. *Advances in Experimental Political Science*.

Vogels, Emily A. , Monica Anderson, Margaret Porteus, Chris Baronavski, Sara Atske, Colleen McClain, Brooke Auxier, Andrew Perrin, and Meera Ramshankar (2021). Americans and 'cancel culture': Where some see calls for accountability, others see censorship, punishment. https://www.pewresearch.org/internet/2021/05/19/americans-and-cancel-culture-where-some-see-calls-for-accountability-others-see-censorship-punishment/.

Vosoughi, Soroush , Deb Roy, and Sinan Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

Vraga, Emily K , Leticia Bode, and Melissa Tully (2022). Creating news literacy messages to enhance expert corrections of misinformation on twitter. *Communication Research 49*(2), 245–267.

Vraga, Emily K and Melissa Tully (2020). News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society 23*(2), 183–199.

Walter, Nathan , Jonathan Cohen, R Lance Holbert, and Yasmin Morag (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication 37*(3), 350–375.

Wardle, Claire and Hossein Derakhshan (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, Volume 27. Council of Europe Strasbourg.

White, Ismail K (2007). When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review 101*(2), 339–354.

Supplementary materials

Appendix A

A-1 Study I

A-1.1 Demographics

Table A.1: Balance check

Statistic	Control	Treatment
% Male	51	45
% Nonwhite	27	29
Age	39 [13.3]	39 [12.9]
Income	$25k - $29k [2.98]	$25k - $29k [3.0]
Ideology	3.8 [1.98]	3.8 [1.91]
% College	60	55
% Support Gun Control	84	85
% Enjoy Pineapples on Pizza	71	67
Political Knowledge	3.0 [1]	3.0 [1]
% Passed Attention Check	92	92

Non-dichotomous values presented with standard deviations in brackets. Gender was coded 0 for females and 1 for males. Nonwhite is 0 for white and 1 for all other races. Income was coded on a 1- 12 scale with 1 being less than $5000 and 12 being $100000 or greater. The meaning of the average in this category is presented (actual value 9) and the standard deviation is on the 1-18 scale. College was coded as 1 if the individual indicated they had an Bachelor's Degree or higher and 0 for any attainment lower. Ideology was a 1-7 measure where higher values represented more conservatism. Pineapple on pizza and gun control positions were binary with 0 indicating positions opposed to pineapples on pizza/gun control and 1 was in favor of pineapples on pizza/gun control. Political knowledge is the average number of correct responses for four questions.

Variable coding

Table A.2: Variable coding scheme

Variable	Description
Congeniality	Preferences for gun control and pineapples on pizza were assessed using two questions each, on a 1-7 scale. Higher values indicate favoring the topic. Binary indicators were created: 0 for opposition and 1 for support. Participants' preferences were indexed, resulting in eight rows per participant in the dataframe, one for each question. See Table 1 for visualization.
Rating	The dependent variable was constructed by subtracting the raw scores of uncongenial arguments from the raw scores of congenial arguments participants reported for each of the eight arguments.
Strength	A binary indicator was constructed to measure strength.
Attention to politics	Attention to politics was a 1-5 scale variable used as a numeric in the regression which asked how often the participant paid attention to news.
Topic	A binary indicator establishing whether the question in the row was about gun control or pineapples on pizza.
Treatment	Treatment was a binary indicator establishing which group the participant.
Strength of partisanship	Strength of partisanship was a follow-up question for those who identified as Republicans or Democrats. The question was binary where 0 was a weak democrat/republican and 1 was a strong democrat/republican.
NTE/NFC	Need for cognition and need to evaluate measures were measured using 2 questions for NFC and 2 questions for NFC. Higher values mean more cognitive/evaluative. The variables were added together and analyzed as numeric values.
Income	The variable was left as a categorical for the analysis.
PID	PID asked participants if they felt closer to the Democrat or Republican party or something else. Coded 0 for Democrat and 1 for Republicans and 2 for all others.
Race	Race listed 7 options for racial composition which included the option to mark multiple boxes. Those who marked multiple boxes were added to a new group, multicultural. The variable was transformed to a binary where 0 represented white and 1 represented all other categories.
Ideology	Ideology was a 1-7 scale where 7 was more conservative and this variable was used as an integer in the analysis.
Political Knowledge	Participants were asked four questions about civics and current events (see pages A-10 and A-12 of Appendix). These questions were put into an additive scale which was then converted to the proportion of correct responses.

A-1.2 Treatment prompt responses

Table A.3: Random writing sample

- Being able to take in all information and points of view allows for compromise and progress to be made. It is the only way to build coalitions and please as many people as possible even though it might be impossible to give exactly what some wanted.
- Because objectivity has to do with logic and logical solutions to problems. Objectivity takes into account all the evidence for or against a problem before coming up with a valid solution.
- It's important to keep an open mind. If you believe something just because your parents believed it then you are taking on their views. You are not thinking for yourself. Change comes when we can keep an open mind and figure things out for ourselves. The old way is not always the best way.
- Because I think it allows things to be looked at in a more fair way. If you approach things subjectively, emotions and feelings get in the way. Being objective and only considering facts instead of individual opinions and feelings works out better overall. Not everyone is going to feel the same about things. We need to make a lot of decisions based on this instead of feelings.
- Objectivity allows an individual to step back and evaluate situations without any emotional biases. It's the basis for individuality and our judicial system.
- Objectivity allows us to make decisions based on facts rather than emotions. In society, this helps us to do what is best for all rather than only what is best for the person making the laws and decisions.
- I think that everyone should be treated as an equal and treated fairly like everyone should be. Everyone's opinion should be listened to and still respect them even if their opinion is different than ours.
- Sometimes, being objective is the only way to ensure fairness to all in the society. For example, I would like my skills to be viewed objectively and compared objectively to other candidates when I am applying for a job, so that the most qualified person gets the position. When people go on trial, it is also important for evidence to be viewed objectively, so that everyone gets a fair trial and due process, without feelings and high emotions to affect the ultimate outcome.
- Objectivity is important to me and society because that is how fair laws are made. The laws impact everyone, not just one group. They are made for the majority not the minority.
- It allows you and others to focus on the facts and address problems that would other wise be unseen through lens of bias or agenda. It also allows for a greater possibility of unity from those of different perspectives otherwise.
- Objectivity is valuable to me and society because you cant be one sided on all situations. We should look at all points of arguments. Based on observation we can make decisions. Not because of who you side with.

A-2 Additional research questions: Political versus non-political

My research builds on the groundwork laid in Lodge and Taber's research and parallels ideas explored in Groenendyk and Krupnikov's. I evaluate the quality of the arguments and (in Study 1) the political nature of the subject. Using a within-subjects design, I estimate the size of the argument congruency effect and directly test whether it is less when the topic of the argument is non-political.

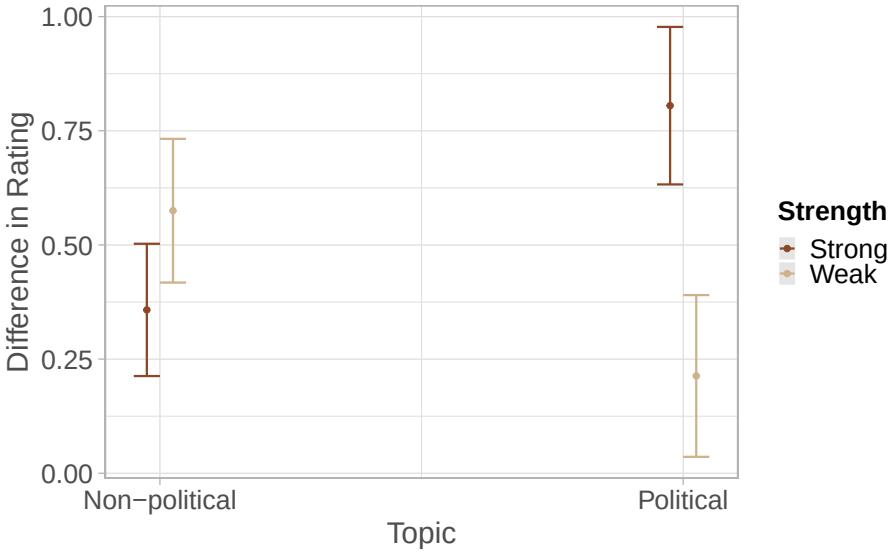
> *Research question 3: Do political topics result in more argument congruency bias than non-political topics?*

The third research question examines whether political topics elicit more argument congruency bias than non-political topics. Figure A.1 compares the level of bias between political and non-political arguments based on their strength.

The analysis reveals mixed results. For strong arguments, argument congruency bias increases by 0.44 points when moving from non-political to political topics, aligning with the expectation that political topics are more emotionally charged and thus more prone to biased evaluations. However, for weak arguments, bias decreases by 0.36 points from non-political to political topics, challenging the assumption that political topics always yield more argument congruency bias.

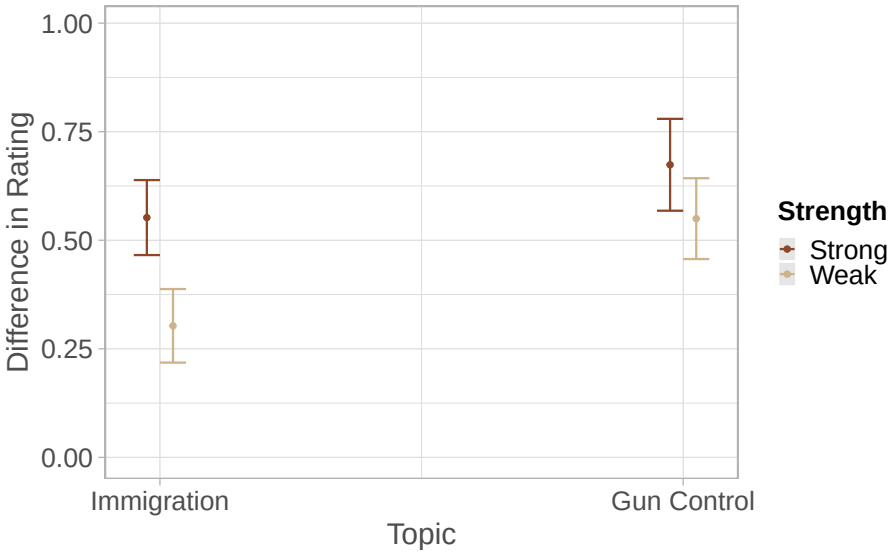
While both the topics in Study II are political, we can still explore any differential effects that each topic may have. When put on the same scale as Figure A.1, we can see that both political topics yield similar levels of argument congruency bias. Weak gun control arguments yield slightly more argument congruency bias than weak immigration arguments.

Figure A.1: Argument congruency bias for strong political arguments and weak non-political arguments



Effect of topic on difference in rating of congenial and uncongenial arguments by strength. Point estimates (with 95% confidence intervals) from OLS regression with clustered standard errors. Non-political refers to arguments about pineapple on pizza, and political refers to arguments about gun control. Control variables include gender, college education, race, ideology, party ID, age, attention to politics, need to evaluate, need for cognition, political knowledge, and strength of partisanship. See Figure A.2 for Study II comparison. See Table A.4 for associated regression analyses.

Figure A.2: Immigration yields slightly less argument congruency bias than gun control



Effect of topic on difference in rating of congenial and uncongenial arguments by strength. Point estimates (with 95% confidence intervals) from OLS regression with clustered standard errors and sample weights. Control variables include gender, college education, race, ideology, party ID, and age. See Figure A.1 for Study I comparison. See ?? for associated regression analyses.

A-2.1 Regressions

Table A.4: Impact of topic and argument strength on argument evaluation in Mechanical Turk sample

	Model 1	Model 2
(Intercept)	0.33	0.49
	(0.32)	(0.30)
Topic	-0.36^{**}	-0.41^{***}
	(0.12)	(0.12)
Strength	-0.22^{*}	-0.25^{*}
	(0.11)	(0.11)
Gender	0.04	0.03
	(0.09)	(0.08)
Education	−0.07	-0.08^{*}
	(0.04)	(0.04)
Nonwhite	−0.04	0.01
	(0.09)	(0.09)
Ideology	−0.04	−0.04
	(0.03)	(0.03)
Republican	−0.06	−0.09
	(0.12)	(0.12)
31-44 years old	0.11	0.13
	(0.10)	(0.09)
45-59 years old	0.08	0.12
	(0.12)	(0.12)
60+ years old	0.20	0.30
	(0.18)	(0.17)
Attention to politics	0.05	0.04
	(0.04)	(0.04)
Need to evaluate	0.10	0.11
	(0.06)	(0.06)
Need for cognition	0.01	−0.00
	(0.03)	(0.03)
Political knowledge	0.11	0.07
	(0.20)	(0.19)
Strength of partisanship	−0.01	−0.04
	(0.09)	(0.08)
Topic×Strength of argument	0.81^{***}	0.85^{***}
	(0.17)	(0.16)
R^2	0.02	0.02
Adj. R^2	0.02	0.02
N.	2201	2383
RMSE	1.95	1.96

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the analysis excluding participants who failed the attention check, while Model 2 includes these participants. Statistical significance is denoted as follows: ***p<0.001, **p<0.01, *p<0.05.

Table A.5: Interaction between argument congruency bias and argument strength in Mechanical Turk sample

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.36	0.47	1.37**	−0.94
	(0.34)	(0.33)	(0.46)	(0.34)
Treatment	−0.29*	−0.28*	−0.28*	−0.34*
	(0.12)	(0.12)	(0.16)	(0.18)
Strength	0.01	−0.01	−0.47**	0.56***
	(0.11)	(0.11)	(0.14)	(0.17)
Gender	0.03	0.03	−0.20	0.27*
	(0.09)	(0.09)	(0.12)	(0.14)
Education	−0.07	−0.08	−0.08	−0.06
	(0.04)	(0.04)	(0.06)	(0.06)
Nonwhite	−0.05	0.01	0.15	−0.29
	(0.10)	(0.10)	(0.13)	(0.15)
Ideology	−0.05	−0.04	0.02	−0.13*
	(0.03)	(0.03)	(0.04)	(0.03)
Republican	−0.05	−0.08	−0.02	−0.08
	(0.13)	(0.13)	(0.18)	(0.18)
31-44 years old	0.11	0.13	0.16	0.07
	(0.10)	(0.10)	(0.13)	(0.16)
45-59 years old	0.08	0.12	0.03	0.13
	(0.13)	(0.13)	(0.16)	(0.19)
60+ years old	0.19	0.29	0.04	0.37
	(0.19)	(0.18)	(0.26)	(0.29)
Attention to politics	0.04	0.03	0.06	0.02
	(0.05)	(0.05)	(0.06)	(0.07)
Need to evaluate	0.10	0.11	−0.01	0.21*
	(0.06)	(0.06)	(0.08)	(0.10)
Need for cognition	0.01	−0.00	−0.02	0.04
	(0.03)	(0.03)	(0.04)	(0.05)
Political knowledge	0.10	0.07	−0.12	0.29
	(0.20)	(0.19)	(0.27)	(0.32)
Strength of partisanship	−0.02	−0.05	−0.13	0.11
	(0.09)	(0.09)	(0.12)	(0.15)
Treatment×Strength of argument	0.34*	0.37*	0.50*	0.07
	(0.16)	(0.15)	(0.21)	(0.22)
R^2	0.02	0.02	0.02	0.07
Adj. R^2	0.01	0.01	0.00	0.06
N.	2201	2383	1123	1078
RMSE	1.96	1.97	1.81	2.05

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the primary analysis excluding participants who failed the attention check. Model 2 includes these participants to assess the robustness of the findings. Model 3 focuses solely on arguments related to pineapples on pizza, isolating the effect within this non-political topic. Model 4 similarly isolates the effect within the political topic of gun control. Statistical significance is denoted as follows: ***p<0.001, **p<0.01, *p<0.05.

A-2.2 Survey instrument

Pre-treatment survey

1. Please select your gender.
    - Male
    - Female
    - Other
2. Please choose one or more race or ethnic group that describes you. Mark all that apply.
    - White non-Hispanic
    - Black or African-American non-Hispanic
    - American Indian or Alaska Native non-Hispanic
    - Hispanic
    - Asian non-Hispanic
    - Native Hawaiian or Other Pacific Islander non-Hispanic
    - Other
3. Please select your year of birth.
    - 2010 ... 1905
4. We want to know about the total income in your household. What was your household income in the past year?
    - Less than $10,000
    - $10,000 - $19,999
    - $20,000 - $29,999
    - $30,000 - $39,999
    - $40,000 - $49,999
    - $50,000 - $59,999
    - $60,000 - $69,999
    - $70,000 - $79,999
    - $80,000 - $89,999
    - $90,000 - $99,999
    - $100,000 - $149,999
    - More than $150,000
5. Are you a U.S. citizen?

- Yes
- No

6. Please indicate the highest level of education you have received.

    - Less than High School
    - High School Graduate
    - Some College but no degree (yet)
    - Associate's Degree
    - Bachelor's Degree
    - Post Graduate Degree (MA, MBA, MD, JD, PhD, etc.)

7. Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what?

    - Republican
    - Democrat
    - Independent
    - Something else

8. With which party do you identify?

9. Would you call yourself a strong Republican or a not very strong Republican?

    - Not very strong Republican
    - Strong Republican

10. Would you call yourself a strong Democrat or a not very strong Democrat?

    - Not very strong Democrat
    - Strong Democrat

11. Do you think of yourself as closer to the Republican or Democratic party?

    - Closer to the Republican party
    - Closer to the Democratic party

12. When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?

    - Very conservative
    - Somewhat conservative
    - Slightly conservative
    - Neither liberal nor conservative; Moderate
    - Slightly liberal
    - Somewhat liberal
    - Very liberal

13. How often do you follow what's going on in government/public affairs?
    - Never
    - Sometimes
    - About half the time
    - Most of the time
    - Always
14. Please indicate how well each statement describes you.
    - I would prefer complex to simple problems.
        - Describes me extremely well
        - Describes me very well
        - Describes me moderately well
        - Describes me slightly well
        - Does not describe me
    - I like to have the responsibility of handling a situation that requires a lot of thinking.
        - Describes me extremely well
        - Describes me very well
        - Describes me moderately well
        - Describes me slightly well
        - Does not describe me
    - I like to have strong opinions. It is important that you pay attention; please select "Describes me slightly well".
        - Describes me extremely well
        - Describes me very well
        - Describes me moderately well
        - Describes me slightly well
        - Does not describe me
    - I have many more opinions than the average person.
        - Describes me extremely well
        - Describes me very well
        - Describes me moderately well
        - Describes me slightly well
        - Does not describe me
    - I often prefer to remain neutral about complex issues.
        - Describes me extremely well
        - Describes me very well
        - Describes me moderately well
        - Describes me slightly well
        - Does not describe me
    - I form opinions about everything.

- Describes me extremely well
- Describes me very well
- Describes me moderately well
- Describes me slightly well
- Does not describe me

15. The following questions will ask questions related to civics and current events in the United States. Please select the answer you believe to be correct.

    - For how many years is a United States Senator elected - that is, how many years are there in one full term of office for a U.S. Senator?
        - Two years
        - Four years
        - Six years
        - Eight years
        - None of these
        - Don't know
    - How many times can an individual be elected President of the United States under current laws?
        - Once
        - Twice
        - Four times
        - Unlimited number of terms
        - Don't know
    - Which party elected the most members to the House of Representatives in the elections in November 2018?
        - Democrats
        - Republicans
        - Neither
    - How many U.S. Senators are there from each state?
        - One
        - Two
        - Four
        - Eight
        - Depends on the state
        - Don't know

Issue positions

16. The following questions will ask about your opinions on two topics: Gun control and pineapples on pizza.

    - Guns like cars should only be used by responsible citizens. Gun control laws just ensure that responsible people are using guns in a responsible manner.

  - Strongly disagree
  - Disagree
  - Somewhat disagree
  - Somewhat agree
  - Agree
  - Strongly agree
- How much do you personally care about the issue of gun control?
  - None at all
  - A little
  - A moderate amount
  - A lot
  - A great deal
- Compared to how you feel about other public issues, how strong are your feelings regarding the issue of gun control?
  - I have stronger feelings about other issues
  - Slightly strong
  - Moderately strong
  - Very strong
  - Extremely strong
- Over the past few years, our right to bear arms has been eroding. This encroachment on our rights must be stopped.
  - Strongly disagree
  - Disagree
  - Somewhat disagree
  - Somewhat agree
  - Agree
  - Strongly agree
- I would not eat a pizza that had pineapple on it.
  - Strongly disagree
  - Disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Agree
  - Strongly agree
- I enjoy eating pineapple on pizza.
  - Strongly disagree
  - Disagree
  - Somewhat disagree
  - Neither agree nor disagree
  - Somewhat agree
  - Agree

  - Strongly agree
- When you eat pizza, how often do you select pineapple as a topping?
  - Never
  - Sometimes
  - About half the time
  - Most of the time
  - Always
  - I do not eat pizza
- Compared to how you feel about other pizza toppings, how strong are your feelings about pineapple on pizza?
  - I do not feel very strongly about pineapple on pizza
  - I have somewhat strong feelings about pineapple on pizza
  - I have strong feelings about pineapple on pizza
  - I have very strong feelings about pineapple on pizza

Treatment

17. Movies can be fun, but don't underestimate how much they can provide to our society. Movies encourage ideas and social commentary within communities. They have the power to express a culture's ideals and shape them. Movies are important because they give us the ability to form lasting human connections by letting us share our experiences with each other.

    - Please write a brief paragraph describing the last movie you saw.

18. Objectivity is the ability to make judgements without relying on personal feelings or personal opinions. Being objective means applying the rules fairly and treating everyone the same rather than showing favoritism. Objectivity requires you to consider perspectives other than your own to achieve less biased conclusions.

    - Please write a brief paragraph explaining why objectivity is valuable to you and society.

19. You will be presented with arguments about two different issues. For each issue, you will read four arguments. Each argument will have a premise and a conclusion; the conclusion advocates a particular position while the premise supplies a reason for the position.

Instructions

For each issue, you will be asked to evaluate the strength of each argument. By 'strength,' we mean the extent to which the conclusion follows from the premise. Thus, your job is to judge the extent to which the conclusion follows from the premise—NOT whether you think the conclusion is true or false.

- REMEMBER: whether you agree or disagree with the conclusion of an argument is not the same thing as the degree to which you think the argument is weak or strong.

Example:

- Strong Argument: "Smoking is linked to cancer and other health issues. Therefore, smoking is bad for your health."
- Weak Argument: "Public establishments ban smoking inside. Therefore, smoking is bad for your health."

20. Crime victims who do not resist are twice as likely to be injured compared to those who defend themselves; reducing access to guns makes self-defense against violent crimes more difficult. Therefore, we do not need more gun control legislation.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

Arguments to be evaluated

21. A study found that you or a member of your family are 43 times more likely to be killed by your own gun than by an intruder's; guns present more risk than protection. Therefore, we need more gun control legislation.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

22. If we allow the government to regulate the use of guns, soon they will be forcibly removing guns from our homes; it is imperative that we prevent involuntary gun removal. Therefore, we do not need more gun control legislation.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

23. Americans have an obsession with guns that is quite disturbing, and gun control opponents are not willing to compromise on the issue. Therefore, we need more gun control legislation.

    - Is this argument weak or strong?

- Very weak
- Moderately weak
- Slightly weak
- Slightly strong
- Moderately strong
- Very strong

24. The juice and sugar from pineapples make the pizza soggy; pineapples ruin the texture of the crust. Therefore, pineapples are a bad topping for pizza.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

25. Pineapples are a good source of vitamin C and dietary fiber; pineapples are a healthy addition to pizza. Therefore, pineapples are a good topping for pizza.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

26. Nobody I have met says they enjoy pineapples on pizza, and I think pineapples are a disgusting topping. Therefore, pineapples are a bad topping for pizza.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

27. Whenever I order pizza with pineapple, my family enjoys it, and I love the taste of pineapples on pizza. Therefore, pineapples are a good topping for pizza.

    - Is this argument weak or strong?
        - Very weak
        - Moderately weak
        - Slightly weak
        - Slightly strong
        - Moderately strong
        - Very strong

A-3 Study II

A-3.1 Demographics

Table A.6: Balance check

Statistic	Control	Treatment
% Male	48	51
% Nonwhite	33	39
Age	47 [17.5]	49 [17.1]
Income	$40k - $49k [4.41]	$25k - $29k [3.0]
Ideology	4.‘ [2.08]	4.1 [2.02]
% College	37	33
% Support Gun Control	78	79
% Enjoy Immigration	67	66

Non-dichotomous values presented with standard deviations in brackets. Gender was coded 0 for females and 1 for males. Nonwhite is 0 for white and 1 for all other races. Income was coded on a 1- 18 scale with 1 being less than $5000 and 18 being $200000 or greater. The meaning of the average in this category is presented (actual value 9) and the standard deviation is on the 1-18 scale. College was coded as 1 if the individual indicated they had an Bachelor's Degree or higher and 0 for any attainment lower. Ideology was a 1-7 measure where higher values represented more conservatism. Pineapple on pizza and gun control positions were binary with 0 indicating positions opposed to immigration/gun control and 1 was in favor of immigration/gun control. Political knowledge is the average number of correct responses for four questions.

A-4 Pre-registration details

Pre-analysis plan

Description of sample

The sample is collected via AmeriSpeak using a NORC National Frame address-based sample. A National Frame Area (NFA) is selected which is made up of a metropolitan area of one or more counties or a county with a population with at least 10,000 citizens. These NFAs are then defined as a Census tract or block group which contains at least 300 housing units according to the 2010 Census. A stratified probability sample of 1,514 segments is selected with probability proportional to size. These addressed based samples are then contacted by US mail, telephone interviewers, overnight express mailers, and face-to-face interviews. For a more detailed explanation, visit this website: https://www.norc.org/PDFs/AmeriSpeak%20Technical%20Overview%202015%2011%2025.pdf.

Overview of design

For the experiment, AmeriSpeak will equally sort individuals into a control and treatment group using computerized randomization to evaluate the effect of *an intervention which emphasizes **objectivity** on individuals' rating of arguments*. The unit of analysis is the difference in rating of arguments they agree with (congenial arguments) minus arguments they do not agree with (uncongenial).

The unit of analysis is the difference in rating of individuals' rating of arguments by strength and topic. The difference in rating of congenial and uncongenial arguments is selected because this indicates how much bias is present when participants encounter arguments which they agree/disagree with. We refer to this as argument congruency bias (ACB).

The intervention is a statement about objectivity and a series of questions which ascertains whether objectivity is an important aspect in their life (see survey instrument for exact wording of the statement and questions).

The main outcome is the difference in argument rating by strength of argument. For example, an individual who is pro-gun control and pro-immigration will have four rows constructed as follows:

1- Strong Pro-Gun Control Argument Rating minus Strong Anti-Gun Control Argument Rating
2- Weak Pro-Gun Control Argument Rating minus Weak Anti-Gun Control Argument Rating
3- Strong Pro-Immigration Argument Rating minus Strong Anti-Immigration Argument Rating
4- Weak Pro-Immigration Argument Rating minus Weak Anti-Immigration Argument Rating

Hypotheses

H1: The treatment will decrease the level of argument congruency bias among treatment participants vs. control participants.

RQ1: Does the treatment differ by topic?

In addition to overall reduction in ACB, we expect that the reduction in ACB will be different depending upon the strength of the argument.

H2: The treatment will be more effective when individuals are evaluating arguments that are weak.

Measures

Dependent variable

The data is transformed so that each individual has four rows, as described above. Prior to the treatment and argument evaluation, we will assess individuals' policy preferences to obtain our dependent variable. The measure of an individuals' policy preference is assessed using two questions per topic.

These questions are in opposite directions (i.e. 1. there should be more gun control 2. There should not be more gun control) therefore, the question against a position (anti-gun control; anti-immigration) will be transformed so that the highest number (4) will be in favor of the topic. Each of these questions will be on a 1-4 scale where one means strongly disagree and four means strongly agree with no option for indifference (exact question wording can be seen in the survey instrument).

The policy position for each topic is then added together and values 5 and above are labeled as Pro-Topic and values 4 or below are labeled as Anti-Topic. We will also look at these variables separately to see if one of the two indicators has more variation than the other. This will entail a separate, but similar analysis. The dependent variable is constructed as defined above; for another illustration consider the opposite scenario where an individual is anti-gun control and anti-immigration:

1- Strong Anti-Gun Control Argument Rating minus Strong Pro-Gun Control Argument Rating
2- Weak Anti-Gun Control Argument Rating minus Weak Pro-Gun Control Argument Rating
3- Strong Anti-Immigration Argument Rating minus Strong Pro -Immigration Argument Rating
4- Weak Anti -Immigration Argument Rating minus Weak Pro -Immigration Argument Rating

By subtracting the argument rating the dependent variable becomes the difference in arguments whose conclusion they like and arguments whose conclusions they don't like. This is a measure of argument congruency bias.

Independent variable

The main explanatory variable is the treatment assignment. We will explore the effects of each topic separately as well by interacting the treatment with a binary indicator for the topic being evaluated. For our second hypothesis, the treatment will be interacted with the strength of the argument. The strength of the argument is constructed by referring to the construction of the DV above. If a particular row for a participant is Strong Anti-Gun Control Argument Rating minus Strong Pro-Gun Control Argument Rating then that row is labeled as strong.

Controls are included for demographic variables (age, race, party ID, gender, ideology) and Single Item Trait measure which ascertains an individuals' level of empathy as this may account for differences in how one evaluates an argument (an empathetic person may be more open to other people's opinions).

Estimation procedure

We estimate the effect of exposure to the concept of objectivity on the level of ACB. Our estimand is the average treatment effect and we will use linear regression to extract this estimate. The broadest question is whether the treatment condition effects levels of bias. This is determined using the following difference in differences.

$$Y = \underbrace{\left[\bar{y}_{(c,T)} - \bar{y}_{(u,T)}\right]}_{\text{Bias in Treatment}} - \underbrace{\left[\bar{y}_{(c,R)} - \bar{y}_{(u,R)}\right]}_{\text{Bias in Control}}$$

Average treatment effect

Where *c* indicates if an argument is congenial and *u* indicates if the argument is uncongenial to the participants' prior beliefs. *T* indicates if an individual is in the treatment condition and *R* indicates an individual is in the control condition.

The next hypothesis askes how the strength of the argument moderates this outcome.

$$Y = \underbrace{\left(\underbrace{\left[\bar{y}_{(c,T,H)} - \bar{y}_{(u,T,H)}\right]}_{\text{Bias in treatment with high quality argument}} - \underbrace{\left[\bar{y}_{(c,R,H)} - \bar{y}_{(u,R,H)}\right]}_{\text{Bias in control with high quality argument}}\right)}_{\text{Average treatment effect among high quality arguments}} - \underbrace{\left(\underbrace{\left[\bar{y}_{(c,T,L)} - \bar{y}_{(u,T,L)}\right]}_{\text{Bias in treatment with low quality argument}} - \underbrace{\left[\bar{y}_{(c,R,H)} - \bar{y}_{(u,R,L)}\right]}_{\text{Bias in control with low quality argument}}\right)}_{\text{Average treatment effect among low quality arguments}}$$

In this equation, *H* indicates arguments of high quality (strong arguments) and *L* indicates arguments of low quality (weak arguments).

We provide an example of the code used in our prior analysis altered to reflect the parameters of our upcoming experiment:

H1: lm_robust(rating ~ treatment + age + strength + gender + income + ideology + PID + empathy, clusters = ID)

RQ1: lm_robust(rating ~ treatment *topic + age + strength + gender + income + ideology + PID+ empathy, clusters = ID)

H2: lm_robust(rating ~ treatment * strength + age + gender + income + ideology + PID+ empathy, clusters = ID)

Inference strategy

We use clustered standard errors at the individual level. For each of the two topics there is a row for strong and weak arguments per individual. Clustered standard errors account for the multiple instances of the same participant. We use a two-tailed test and we will reject the null when the p-value is less than .05.

The main results will be between-subject and will be pooled so that both topics are included in the estimation. We will also examine how the results vary by topic. The second hypothesis will be examined using the same strategy except in this portion, we will look within-subjects and observe how strength interacts with the treatment effect.

Data issues

One possible challenge in our study is the possibility of individuals skipping questions. AmeriSpeak requires that individuals have the option to skip questions and this could pose problems at multiple points. For one, if individuals choose not to answer questions about their policy positions, then we will not be able to construct the dependent variable as it necessarily relates to their prior beliefs on issues. Further, if participants skip the questions which ask them to evaluate arguments then we will not be able to construct the dependent variable.

Given the structure of AmeriSpeak there is no way to avoid the possibility of missing values, where in the past we were able to force responses to questions. If missingness occurs in the dependent variable, we will be forced to assume that these values are missing at random and exclude them from the analysis. If for any covariate, more than 30% of the data is missing will choose to exclude that measure from our control variables.

Survey instrument

Please indicate your support for the following statements.

MONTLOCKPOLICY01S3
Guns, like cars, should only be used by responsible citizens. Gun control laws just ensure that responsible people are using guns in a responsible manner.
[RESPONSE OPTIONS]

1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree

MONTLOCKPOLICY02S3

A-3.2 Regressions

Table A.7: Interaction between argument congruency bias and argument strength in The American Social Survey sample

	Main Analysis	Immigration only	Gun control only
(Intercept)	0.38	0.61	0.16
	(0.25)	(0.31)	(0.37)
Treatment	0.00	−0.04	0.04
	(0.10)	(0.13)	(0.13)
Strength	0.21**	0.19*	0.21*
	(0.07)	(0.09)	(0.08)
Gender	−0.05	0.03	−0.14
	(0.08)	(0.10)	(0.12)
Education	0.07	−0.05	0.19*
	(0.05)	(0.06)	(0.08)
Nonwhite	0.17*	0.03	0.32*
	(0.08)	(0.10)	(0.12)
Ideology	−0.01	−0.03	0.01
	(0.02)	(0.03)	(0.04)
Moderate Democrat	−0.32**	−0.44**	−0.20
	(0.11)	(0.14)	(0.16)
Lean Democrat	−0.16	−0.24	−0.06
	(0.12)	(0.16)	(0.17)
Don't Lean/independent/None	−0.36*	−0.24	−0.47*
	(0.15)	(0.17)	(0.20)
Lean Republican	−0.60***	−0.46*	−0.71***
	(0.13)	(0.19)	(0.19)
Moderate Republican	−0.27	0.13	−0.65*
	(0.16)	(0.17)	(0.28)
Strong Republican	−0.46**	0.03	−0.93**
	(0.17)	(0.22)	(0.27)
30-44 years old	0.16	0.10	0.23
	(0.12)	(0.15)	(0.17)
45-59 years old	0.19	0.13	0.25
	(0.13)	(0.16)	(0.21)
60+ years old	0.23	0.04	0.43*
	(0.12)	(0.15)	(0.19)
Treatment×Strength of argument	−0.11	−0.01	−0.22
	(0.10)	(0.14)	(0.13)
R^2	0.03	0.03	0.07
Adj. R^2	0.02	0.02	0.07
N.	3861	1920	1941
RMSE	1.48	1.34	1.56

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the primary analysis. Model 2 isolates the model to the topic of immigration only and model 3 focuses solely on gun control. Statistical significance is denoted as follows: ***p<0.001, **p<0.01, *p<0.05.

A-4.1 Survey instrument

Issue positions

1. The following questions will ask about your opinions on two topics: Gun control and immigration.
    - There should be no limits on the number of guns someone can own. It is not the government's job to pick and choose the types of weapons it finds acceptable for citizens to own.
        - Strongly disagree
        - Disagree
        - Somewhat disagree
        - Somewhat agree
        - Agree
        - Strongly agree
    - America is a nation built by immigrants; therefore, we ought to continue to allow immigrants in to enrich this nation and contribute to its diversity
        - Strongly disagree
        - Disagree
        - Somewhat disagree
        - Somewhat agree
        - Agree
        - Strongly agree
    - Illegal immigration hurts American workers; burdens American taxpayers; undermines public safety' and places enormous strain on local schools, hospitals, and communities in general, taking precious resources away from the poorest American who need them most.
        - Strongly disagree
        - Disagree
        - Somewhat disagree
        - Somewhat agree
        - Agree
        - Strongly agree
    - Please indicate how well the following statement describes you:
      I am an empathetic person
        - 1- Not very true of me
        - 2
        - 3
        - 4
        - 5 - Very true of me

Treatment questions

1. Objectivity is important because it allows people to think carefully about opinions that differ from their own.
    - Strongly disagree
    - Disagree
    - Somewhat disagree
    - Somewhat agree
    - Agree
    - Strongly agree
2. Being objective makes it easier for people to get as close to the truth as possible.
    - Strongly disagree
    - Disagree
    - Somewhat disagree
    - Somewhat agree
    - Agree
    - Strongly agree
3. Objectivity is important to society as it ensures that rules are applied fairly.
    - Strongly disagree
    - Disagree
    - Somewhat disagree
    - Somewhat agree
    - Agree
    - Strongly agree
4. Objectivity is important because it allows you to examine facts without letting emotions get in the way.
    - Strongly disagree
    - Disagree
    - Somewhat disagree
    - Somewhat agree
    - Agree
    - Strongly agree

Instructions

You will be presented with arguments about gun control and immigration. Each argument will have a premise and a conclusion; the conclusion advocates a particular position, while the premise supplies a reason for the position.

An example of a premise is: Working parents should be able to see their children during the day. An example of a conclusion: Therefore, all companies with over 200 employees should provide day care.

For each issue, you will be asked to evaluate the strength of each argument. By 'strength,' we mean the extent to which the conclusion follows from the premise. Thus, your job is to judge the extent to which the conclusion follows from the premise—not whether you think the premise or conclusion is true or false.

Remember: whether you agree or disagree with the premise or conclusion of an argument is not the same thing as the degree to which you think the argument is weak or strong.

Example:

- **Strong Argument:** Smoking is linked to cancer and other health issues. Therefore, smoking is bad for your health.
- **Weak Argument:** Public establishments ban smoking inside. Therefore, smoking is bad for your health.

Arguments to be Evaluated

In this section, you will be asked to read and evaluate eight arguments.

A-4.1.1 Gun Control Arguments

1. **Argument 1:** A main reason why our murder rate is so high is that most crime victims do not resist. According to a prominent study, victims are twice as likely to be injured compared to those who defend themselves. Carrying a gun is thus one's ultimate protection against violent crime. Therefore, gun control legislation should not be implemented.
2. **Argument 2:** A study in a prominent medical journal found that you or a member of your family are 43 times more likely to be killed by your own gun than by an intruder's. Guns aren't the protection many people think they are. Therefore, stricter gun control legislation should be implemented.
3. **Argument 3:** Americans love guns and no true American could possibly support stricter gun control. Therefore, gun control legislation should not be implemented.
4. **Argument 4:** America's obsession with guns is disturbing. Therefore, stricter gun control legislation should be implemented.

A-4.1.2 Immigration Arguments

1. **Argument 5:** According to a reputable immigration research organization, the average household headed by an immigrant receives 41 percent more in federal welfare than the average American

household. Therefore, legislation to reduce immigration should be enacted to decrease the welfare state.

2. **Argument 6:** A study in a reputable public policy organization found that the output in the economy is higher and grows faster with more immigrants. Therefore, legislation to reduce immigration should not be enacted as to facilitate more economic growth.
3. **Argument 7:** Some immigrants are criminals. Therefore, legislation to reduce immigration should be enacted to lower crime.
4. **Argument 8:** All the immigrants I know are great people. Therefore, legislation to reduce immigration should not be enacted.

Appendix B

B-1 Pre-registration details

Title: The Impact of Source Credibility on Misinformation Correction among Black and Latino Americans

Research Question: To what extent can culturally relevant sources more effectively reduce beliefs in misinformation among Black and Latino Americans compared to generic sources?

Experimental Design: We conducted two survey experiments examining the effectiveness of corrections to targeted misinformation from different sources. The experiments used social media posts containing false claims that were widely circulated during the 2020 U.S. election to depress turnout among Latino (Experiment 1) and Black (Experiment 2) voters.

In Experiment 1, participants were shown an image resembling a Facebook post suggesting that ICE agents were arresting people at polling stations. Experiment 2 used a similar post suggesting local police were arresting individuals with outstanding warrants or parking tickets at polling locations.

Participants in each experiment were randomly assigned to one of four conditions:

1. Control: No social media post shown
2. No Correction: Post with misinformation, no correction
3. Generic Correction: Post with misinformation, correction from the ACLU
4. Culturally Relevant Correction: Post with misinformation, correction from UnidosUS (Experiment 1) or NAACP (Experiment 2)

In conditions 3 and 4, corrections were delivered via comments on the posts from the respective organizations.

Dependent Variable: Our primary dependent variable is belief in the specific misinformation claims (misinformation belief). We measured this both before (pre) and after (post) exposure to the experimental treatments using the following questions:

- Experiment 1: "Some have claimed that immigration officials arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?"

- Experiment 2: "Some have claimed that local police arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?"

Response options were: Very confident, somewhat confident, not too confident, not at all confident. We examine the change in misinformation belief from pre to post.

Hypotheses:

1. *Corrective comments will reduce misinformation belief relative to the no correction and control conditions.*
2. *Corrections from culturally relevant organizations will more effectively reduce misinformation belief:*
    - For Latino respondents, reductions will be greater for corrections from UnidosUS vs. the ACLU.
    - For Black respondents, reductions will be greater for corrections from the NAACP vs. the ACLU.

Research Questions:

1. Will pooling Latino and Black respondents reveal greater responsiveness to in-group vs. out-group organizations?
2. How will white respondents respond to corrections from Black and Latino advocacy organizations compared to a generic source?

Analysis Plan: We will compare the average within-subject change in misinformation belief from pre to post between experimental conditions, both overall and within racial/ethnic subgroups. To assess our research questions, we will pool data across experiments and create indicators for culturally relevant source cues.

Sample and Data: We expect a sample of approximately 1750 participants, including 500 Black and 500 Latino respondents, from the "American Social Survey" conducted by the Weidenbaum Center via NORC.

Pilot data (n=52) was collected in early November 2022 and examined only to confirm survey functionality. The full data was collected in late November and early December 2022, remaining inaccessible to researchers until after this pre-registration.

B-2 Study I

B-2.1 Survey instrument

1. How closely do you follow national politics?
    - Very Closely
    - Closely
    - Somewhat Closely
    - Not Closely
2. How closely do you follow politics in your community?
    - Very Closely
    - Closely
    - Somewhat Closely
    - Not Closely
3. Have you ever used a website (including Craigslist, Zillow, and others) to search for housing to rent or own?
    - Yes
    - No
4. Do you favor or oppose a law making English the official language of the United States, meaning most government business would be conducted in English only?
    - Favor
    - Oppose
    - Neither favor, nor oppose
5. How important is being able to speak English in making someone a true American?
    - Very Important
    - Somewhat Important
    - Somewhat Unimportant
    - Very Unimportant

Misperceptions pre-measure

6. Some have claimed that immigration officials arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?

    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident

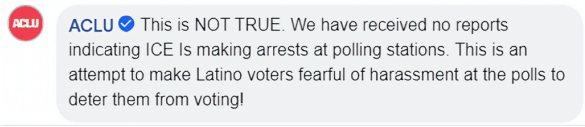
7. Some have claimed that local police arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?

    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident

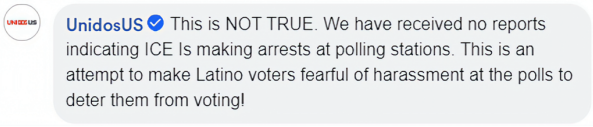
B-2.2 Intervention

Figure B.1: Experiment 1 treatment conditions

(a) Facebook post with no correction



(b) Generic reply

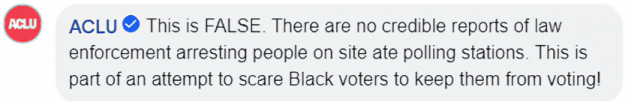


(c) Culturally relevant reply

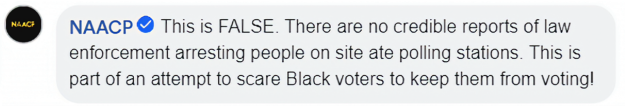
The control condition saw neither the post nor the correction and still received the pre-/post-treatment beliefs in these misperceptions

Figure B.2: Experiment 2 treatment conditions

(a) Facebook post with no correction



(b) Generic reply



(c) Culturally relevant reply

Misperceptions post-measure

8. Some have claimed that local police arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?

    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident

B-2.3 Regressions

Table B.1: Effect of exposure to culturally-relevant correction on misperceptions among Latino participants

	Model 1	Model 2
(Intercept)	1.99***	-0.05
	(0.04)	(0.04)
Latino	0.10	0.09
	(0.09)	(0.08)
No correction	-0.15*	-0.05
	(0.06)	(0.05)
General correction	-0.15*	-0.09
	(0.06)	(0.05)
Cultural correction	-0.19***	-0.12*
	(0.06)	(0.05)
Latino × No correction	-0.04	0.01
	(0.12)	(0.11)
Latino × General correction	0.01	0.11
	(0.12)	(0.11)
Latino × Culturally relevant correction	-0.03	-0.09
	(0.12)	(0.12)
R^2	0.01	0.01
N Respondents	1965	1965

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the treatments impact on misperceptions levels (within-subject experiment), while Model 2 explores an alternative construction of misperceptions where the post-treatment misperceptions were subtracted from the pre-treatment values. ***p<0.001, **p<0.01, *p<0.05.

Table B.2: Effect of exposure to culturally-relevant correction on misperceptions among Black participants

	Model 1	Model 2
(Intercept)	1.90***	-0.04
	(0.04)	(0.03)
Black	0.35***	0.02
	(0.08)	(0.07)
No correction	-0.08	0.02
	(0.06)	(0.05)
General correction	-0.16**	-0.15**
	(0.06)	(0.05)
Cultural correction	-0.25***	-0.23***
	(0.05)	(0.05)
Black × No correction	0.12	0.03
	(0.13)	(0.12)
Black × General correction	0.02	0.06
	(0.13)	(0.12)
Black × Cultural correction	-0.01	-0.05
	(0.12)	(0.12)
R^2	0.06	0.02
N Respondents	1965	1964

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the treatments impact on misperceptions levels (within-subject experiment), while Model 2 explores an alternative construction of misperceptions where the post-treatment misperceptions were subtracted from the pre-treatment values. ***p<0.001, **p<0.01, *p<0.05.

Table B.3: Effect of exposure to culturally-relevant correction on misperceptions among White participants only

	Model 1	Model 2
Constant	1.82***	1.78***
	(0.05)	(0.05)
No correction	−0.10	−0.01
	(0.07)	(0.07)
Genereal correction	−0.10	−0.14*
	(0.07)	(0.07)
Cultural correction	−0.15*	−0.22**
	(0.07)	(0.07)
R^2	0.01	0.02
N	896	896

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the findings of the Latino experiment ant Model 2 displays the Black experiment results among White participants only.

B-3 Study II

B-3.1 Survey instrument

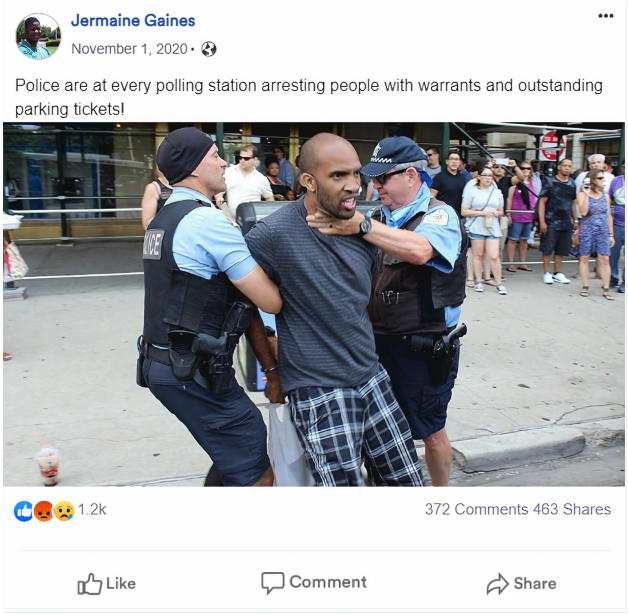
Misperceptions pre-measure

1. Some have claimed that immigration officials arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?
    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident
2. Some have claimed that local police arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?
    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident

Misperceptions post-measure

3. Some have claimed that local police arrested people at the polls during the 2020 election. To the best of your knowledge, how confident are you that this is accurate?
    - Very confident
    - Somewhat confident
    - Not too confident
    - Not at all confident

Figure B.3: Study II treatment conditions



(a) Facebook post with no correction

(b) Black candidate correction

(c) White candidate correction

B-3.2 Regressions

Table B.4: Effect of exposure to culturally-relevant correction on misperceptions

	Model 1	Model 2
(Intercept)	1.87***	-0.03
	(0.05)	(0.06)
Black correction	-0.16*	-0.30***
	(0.06)	(0.06)
White correction	-0.13	-0.18**
	(0.09)	(0.06)
Black	0.53***	0.07
	(0.08)	(0.08)
Black × black candidate	0.00	0.17
	(0.11)	(0.11)
Black × white candidate	-0.13	-0.01
	(0.12)	(0.11)
R^2	0.05	0.04
N Respondents	1422	1422

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Model 1 presents the treatments impact on misperceptions levels (within-subject experiment), while Model 2 explores an alternative construction of misperceptions where the post-treatment misperceptions were subtracted from the pre-treatment values. ***p<0.001, **p<0.01, *p<0.05.

Appendix C

C-1 Study I

C-1.1 Demographics

Table C.1: Balance check

Group	% Male	Age	PID	Income	Ideology
Control	50.3	40	2.2	$30,000–$39,000	2.048
Treatment 1	46.6	40	2.13	$30,000–$39,000	2.021
Treatment 2	47.2	40	2.19	$30,000–$39,000	2.027
Treatment 3	47.0	40	2.23	$30,000–$39,000	2.121
Treatment 4	47.0	40	2.22	$30,000–$39,000	2.073

Note: Party Identification (PID) on a scale from 1 (strong Democrat) to 4 (strong Republican). Ideology is on a 1 to 3 scale where 1 is Liberal, 2 is Moderate, and 3 is Conservative.

Table C.2: Summary of variables

Variable	Description
Treatment groups	Participants were assigned to one of five groups to assess the influence of different types of responses to a racially charged tweet: Control, Tweet-Only, Normative Replies, Counter-Normative Replies, and Mixed Replies.
Opinions	Measured by participants' responses to six statements about the appropriateness of the announcer's language, rated on a five-point Likert scale. Responses were aggregated into a composite score normalized between 0 and 1. See page 172 for the exact phrasing.
F.I.R.E. battery	Assesses fear, acknowledgment of institutionalized racism, and empathy. Participants rated their agreement with four statements on a four-point scale, with higher scores indicating more agreement with the statements. The institutional question and racism question were each recoded so that higher values mean more racist attitudes, to align with all other interpretations in this analysis. See page 166 for exact wording.
Racial resentment scale	Widely used in political science to assess anti-Black affect, beliefs about work ethic, and denial of discrimination. Presented both before and after treatment to measure changes in attitudes. The four questions were added up and compressed down to a zero 1 scale. The median level of racial resentment was .583 those with values below this number were ranked as having low racial resentment and all above ranked as having high racial resentment. See page 166 for exact wording.
Dehumanization scale	Higher values means more human. Racism detected by subtracting rating of Black people from the rating of White people resulting in values greater than 0 representing more racism
Date	Participants were asked if they thought it was all right for white and black people to date. Variable was recoded so that higher values mean more racist attitudes.
Marry	Participants were asked to rate whether they would be comfortable with different groups marrying a family member. Measure created by subtracting feelings about a family marrying a white individual minus a black individual. See page 173 for exact wording.

C-1.2 Regressions

Table C.3: Impact of post/comments on F.I.R.E battery among White participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.89^{***}	2.00^{***}	1.96^{***}	1.56^{***}
	(0.10)	(0.11)	(0.11)	(0.09)
Tweet only	-0.07	0.03	-0.04	0.00
	(0.07)	(0.07)	(0.07)	(0.05)
Normative comments	-0.07	-0.00	-0.01	-0.04
	(0.07)	(0.07)	(0.07)	(0.05)
Mixed comments	-0.07	-0.03	-0.09	0.01
	(0.07)	(0.07)	(0.07)	(0.05)
Counter-normative comments	-0.05	0.03	-0.07	0.01
	(0.07)	(0.07)	(0.07)	(0.05)
Bachelor's	0.36^{***}	-0.25^{***}	0.29^{***}	0.11^{*}
	(0.05)	(0.06)	(0.06)	(0.04)
30 – 44 years old	0.21^{**}	0.12	0.30^{***}	-0.01
	(0.07)	(0.08)	(0.08)	(0.06)
45 – 49 years old	0.26^{***}	-0.04	0.21^{**}	-0.04
	(0.07)	(0.07)	(0.07)	(0.06)
60+ years old	0.28^{***}	-0.14	0.15	-0.07
	(0.07)	(0.08)	(0.08)	(0.06)
Male	0.22^{***}	0.13^{**}	0.35^{***}	0.24^{***}
	(0.04)	(0.04)	(0.05)	(0.03)
Income	-0.10^{***}	0.02	-0.05^{***}	-0.02^{**}
	(0.01)	(0.01)	(0.01)	(0.01)
R^2	0.06	0.02	0.05	0.03
N	2314	2209	2247	2249

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Responses to F.I.R.E battery questions. Includes White participants that passed the attention check. F.I.R.E model was constructed to look at four underlying sources of racism (separately). Regression models predicting racial attitudes based on their treatment condition. Model 1 (Fear) uses fear of other races as the dependent variable. Model 2 (Institutional) uses belief that racial problems are rare and isolated as the dependent variable. Model 3 (Rare) uses the belief that racial problems in the U.S. are rare and isolated as the dependent variable. Model 4 (Empathy) uses respondents' self-reported racial empathy as the dependent variable.

Table C.4: Impact of post/comments and racial resentment level on F.I.R.E battery among White participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.60***	1.49***	1.38***	1.34***
	(0.11)	(0.09)	(0.10)	(0.09)
Tweet only	−0.09	−0.08	−0.15	−0.06
	(0.09)	(0.06)	(0.08)	(0.06)
Normative comments	−0.03	−0.04	0.04	−0.06
	(0.08)	(0.06)	(0.08)	(0.06)
Mixed comments	−0.11	−0.05	−0.08	−0.03
	(0.08)	(0.06)	(0.08)	(0.06)
Counter-normative comments	−0.02	0.07	−0.09	0.02
	(0.08)	(0.07)	(0.08)	(0.06)
High racial resentment	0.56***	0.99***	1.17***	0.40***
	(0.09)	(0.09)	(0.08)	(0.07)
Bachelor's	0.35***	−0.24***	0.31***	0.12**
	(0.05)	(0.05)	(0.05)	(0.04)
30 – 44 years old	0.19**	0.08	0.24***	−0.02
	(0.07)	(0.06)	(0.07)	(0.06)
45 – 49 years old	0.28***	−0.01	0.23***	−0.02
	(0.07)	(0.06)	(0.06)	(0.06)
60+ years old	0.33***	−0.08	0.24***	−0.02
	(0.07)	(0.06)	(0.07)	(0.06)
Male	0.17***	0.07	0.27***	0.21***
	(0.04)	(0.04)	(0.04)	(0.03)
Income	−0.10***	0.03**	−0.05***	−0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Tweet only × high racial resentment	0.01	0.14	0.13	0.09
	(0.13)	(0.12)	(0.12)	(0.11)
Normative comments × high racial resentment	−0.06	0.11	−0.04	0.06
	(0.13)	(0.12)	(0.12)	(0.10)
Mixed comments × high racial resentment	0.08	0.05	0.01	0.08
	(0.13)	(0.12)	(0.12)	(0.10)
Counter-normative comments × high racial resentment	−0.05	−0.08	0.03	0.01
	(0.13)	(0.12)	(0.12)	(0.11)
R^2	0.13	0.29	0.35	0.10
N	2287	2195	2229	2222

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes White participants that passed the attention check. Regression models predicting racial attitudes based on their treatment condition and level of racial resentment. In all instances, higher values indicate more anti-black racism. Model 1 (Fear) uses fear of other races as the dependent variable. Model 2 (Institutional) uses belief that racial problems are rare and isolated as the dependent variable. Model 3 (Rare) uses the belief that racial problems in the U.S. are rare and isolated as the dependent variable. Model 4 (Empathy) uses respondents' self-reported racial empathy as the dependent variable.

Table C.5: Impact of post/comments on F.I.R.E battery among all participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.82^{***}	2.07^{***}	1.96^{***}	1.58^{***}
	(0.09)	(0.10)	(0.10)	(0.08)
Tweet only	-0.06	-0.07	-0.06	-0.01
	(0.06)	(0.06)	(0.06)	(0.05)
Normative comments	-0.04	-0.02	-0.02	-0.05
	(0.06)	(0.06)	(0.06)	(0.05)
Mixed comments	-0.08	-0.07	-0.11	0.00
	(0.06)	(0.06)	(0.06)	(0.05)
Counter-normative comments	-0.02	-0.01	-0.05	-0.01
	(0.06)	(0.06)	(0.06)	(0.05)
Bachelor's	0.39^{***}	-0.22^{***}	0.32^{***}	0.10^{*}
	(0.05)	(0.05)	(0.05)	(0.04)
30 – 44 years old	0.24^{***}	0.08	0.30^{***}	-0.01
	(0.07)	(0.07)	(0.08)	(0.06)
45 – 49 years old	0.27^{***}	-0.04	0.20^{**}	-0.03
	(0.06)	(0.06)	(0.07)	(0.05)
60+ years old	0.26^{***}	-0.14^{*}	0.10	-0.07
	(0.07)	(0.07)	(0.07)	(0.06)
Male	0.24^{***}	0.13^{***}	0.41^{***}	0.25^{***}
	(0.04)	(0.04)	(0.04)	(0.03)
Income	-0.10^{***}	0.01	-0.06^{***}	-0.03^{**}
	(0.01)	(0.01)	(0.01)	(0.01)
Nonwhite	0.29^{***}	-0.27^{***}	-0.05	-0.06
	(0.05)	(0.04)	(0.05)	(0.04)
R^2	0.08	0.03	0.06	0.03
N	2953	2857	2894	2887

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes all participants that passed the attention check. Regression models predicting racial attitudes based on their treatment condition and level of racial resentment. In all instances, higher values indicate more anti-black racism. Model 1 (Fear) uses fear of other races as the dependent variable. Model 2 (Institutional) uses belief that racial problems are rare and isolated as the dependent variable. Model 3 (Rare) uses the belief that racial problems in the U.S. are rare and isolated as the dependent variable. Model 4 (Empathy) uses respondents' self-reported racial empathy as the dependent variable.

Table C.6: Impact of post/comments and racial resentment level on F.I.R.E battery among all participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.55***	1.64***	1.43***	1.34***
	(0.10)	(0.08)	(0.09)	(0.08)
Tweet only	−0.06	−0.13*	−0.13	−0.02
	(0.08)	(0.05)	(0.07)	(0.05)
Normative comments	−0.00	−0.05	0.02	−0.04
	(0.08)	(0.06)	(0.07)	(0.05)
Mixed comments	−0.09	−0.08	−0.09	−0.00
	(0.08)	(0.05)	(0.07)	(0.05)
Counter-normative comments	−0.02	0.05	−0.09	0.02
	(0.08)	(0.06)	(0.07)	(0.05)
High racial resentment	0.55***	0.91***	1.12***	0.43***
	(0.08)	(0.08)	(0.08)	(0.07)
Bachelor's	0.38***	−0.23***	0.31***	0.10**
	(0.05)	(0.04)	(0.04)	(0.04)
30 – 44 years old	0.22**	0.05	0.24***	−0.01
	(0.07)	(0.06)	(0.06)	(0.06)
45 – 49 years old	0.29***	−0.03	0.21***	−0.02
	(0.06)	(0.05)	(0.06)	(0.05)
60+ years old	0.32***	−0.07	0.20**	−0.02
	(0.07)	(0.06)	(0.06)	(0.05)
Male	0.18***	0.05	0.30***	0.20***
	(0.04)	(0.03)	(0.04)	(0.03)
Income	−0.09***	0.02*	−0.05***	−0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Nonwhite	0.34***	−0.19***	0.05	−0.03
	(0.05)	(0.04)	(0.04)	(0.03)
Tweet only × high racial resentment	−0.01	0.09	0.10	0.01
	(0.12)	(0.10)	(0.11)	(0.09)
Normative comments × high racial resentment	−0.05	0.11	−0.01	0.02
	(0.12)	(0.10)	(0.11)	(0.09)
Mixed comments × high racial resentment	0.03	0.07	0.02	0.03
	(0.12)	(0.10)	(0.11)	(0.10)
Counter-normative comments × high racial resentment	−0.02	−0.13	0.05	−0.05
	(0.12)	(0.11)	(0.11)	(0.09)
R^2	0.14	0.26	0.33	0.10
N	2919	2841	2871	2855

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes all particpants that passed the attention checks. Regression models predicting racial attitudes based on their treatment condition and level of racial resentment. In all instances, higher values indicate more anti-black racism. Model 1 (Fear) uses fear of other races as the dependent variable. Model 2 (Institutional) uses belief that racial problems are rare and isolated as the dependent variable. Model 3 (Rare) uses the belief that racial problems in the U.S. are rare and isolated as the dependent variable. Model 4 (Empathy) uses respondents' self-reported racial empathy as the dependent variable.

Table C.7: Impact of post/comments on attitudes and opinions of White participants

	Model 1	Model 2	Model 3	Model 4
Constant	0.05	8.06***	20.89***	0.46***
	(0.06)	(2.08)	(3.06)	(0.02)
Tweet only	−0.08*	−1.65	−0.60	−0.02*
	(0.04)	(1.34)	(1.81)	(0.01)
Normative comments	−0.05	−1.43	−3.25	−0.01
	(0.04)	(1.27)	(1.71)	(0.01)
Mixed comments	−0.00	−1.00	0.66	−0.01
	(0.04)	(1.32)	(1.71)	(0.01)
Counter-normative comments	−0.03	−0.36	−0.43	−0.00
	(0.04)	(1.36)	(1.75)	(0.01)
Bachelor's	−0.01	4.93***	3.13*	0.04***
	(0.03)	(0.98)	(1.48)	(0.01)
30 – 44 years old	0.00	0.44	−1.54	0.05***
	(0.04)	(1.47)	(2.28)	(0.01)
45 – 49 years old	−0.04	2.57	−2.44	0.04***
	(0.04)	(1.39)	(2.06)	(0.01)
60+ years old	−0.09*	3.75*	−1.89	0.04**
	(0.04)	(1.55)	(2.22)	(0.01)
Male	0.01	3.00***	3.09**	0.07***
	(0.02)	(0.85)	(1.15)	(0.01)
Income	−0.00	−0.94***	−0.97**	−0.01***
	(0.01)	(0.23)	(0.32)	(0.00)
R^2	0.01	0.02	0.01	0.06
N	2107	2612	2657	2699

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

White participants only. Regression models predicting racial attitudes based on the treatment condition. Model 1 (Date) uses the belief that it is acceptable for whites and blacks to date each other as the dependent variable. Model 2 (Dehumanization) uses the dehumanization scale as the dependent variable. Model 3 (Marry) uses respondents' reported comfort with a close relative marrying a black person as the dependent variable. Model 4 (Announcer opinions) uses a scale constructed from 4 items measuring opinions regarding the announcer as the dependent variable.

Table C.8: Impact of post/comments and racial resentment level on attitudes and opinions of White participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.54***	3.18	13.71***	0.38***
	(0.08)	(2.22)	(3.14)	(0.02)
Tweet only	−0.11	−2.49	−3.51	−0.03*
	(0.05)	(1.66)	(2.13)	(0.01)
Normative comments	−0.08	0.90	−1.93	−0.00
	(0.05)	(1.62)	(2.09)	(0.01)
Mixed comments	−0.04	−0.00	−0.52	−0.01
	(0.06)	(1.63)	(2.08)	(0.01)
Counter-normative comments	−0.08	−0.13	0.28	0.01
	(0.05)	(1.62)	(2.16)	(0.01)
High racial resentment	0.33***	9.32***	13.18***	0.16***
	(0.07)	(1.86)	(2.31)	(0.01)
Bachelor's	0.07	4.79***	3.00*	0.03***
	(0.04)	(0.97)	(1.46)	(0.01)
30 – 44 years old	−0.13*	0.04	−1.97	0.05***
	(0.06)	(1.47)	(2.22)	(0.01)
45 – 49 years old	−0.15**	2.74*	−1.79	0.04***
	(0.05)	(1.40)	(2.01)	(0.01)
60+ years old	−0.11	4.45**	−0.60	0.05***
	(0.06)	(1.55)	(2.16)	(0.01)
Male	0.17***	2.39**	2.09	0.06***
	(0.03)	(0.85)	(1.14)	(0.01)
Income	−0.04***	−0.83***	−0.83**	−0.01***
	(0.01)	(0.23)	(0.32)	(0.00)
Tweet only × high racial resentment	0.09	0.74	4.51	−0.00
	(0.09)	(2.64)	(3.50)	(0.02)
Normative comments × high racial resentment	−0.03	−4.76	−2.61	−0.02
	(0.09)	(2.53)	(3.40)	(0.02)
Mixed comments × high racial resentment	0.03	−2.04	2.23	−0.00
	(0.10)	(2.62)	(3.36)	(0.02)
Counter-normative comments × high racial resentment	0.04	−0.46	−1.17	−0.02
	(0.09)	(2.70)	(3.48)	(0.02)
R^2	0.09	0.06	0.07	0.24
N	2275	2564	2609	2663

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes White participants that passed the attention check. Regression models predicting racial attitudes based on their treatment condition and level of racial resentment. Model 1 (Date) uses the belief that it is acceptable for whites and blacks to date each other as the dependent variable. Model 2 (Dehumanization) uses the dehumanization scale as the dependent variable. Model 3 (Marry) uses respondents' reported comfort with a close relative marrying a black person as the dependent variable. Model 4 (Announcer opinions) uses a scale constructed from 4 items measuring opinions regarding the announcer as the dependent variable.

Table C.9: Impact of post/comments level on attitudes and opinions of all participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.82***	2.07***	1.96***	1.58***
	(0.09)	(0.10)	(0.10)	(0.08)
Tweet only	−0.06	−0.07	−0.06	−0.01
	(0.06)	(0.06)	(0.06)	(0.05)
Normative comments	−0.04	−0.02	−0.02	−0.05
	(0.06)	(0.06)	(0.06)	(0.05)
Mixed comments	−0.08	−0.07	−0.11	0.00
	(0.06)	(0.06)	(0.06)	(0.05)
Counter-normative comments	−0.02	−0.01	−0.05	−0.01
	(0.06)	(0.06)	(0.06)	(0.05)
Bachelor's	0.39***	−0.22***	0.32***	0.10*
	(0.05)	(0.05)	(0.05)	(0.04)
30 – 44 years old	0.24***	0.08	0.30***	−0.01
	(0.07)	(0.07)	(0.08)	(0.06)
45 – 49 years old	0.27***	−0.04	0.20**	−0.03
	(0.06)	(0.06)	(0.07)	(0.05)
60+ years old	0.26***	−0.14*	0.10	−0.07
	(0.07)	(0.07)	(0.07)	(0.06)
Male	0.24***	0.13***	0.41***	0.25***
	(0.04)	(0.04)	(0.04)	(0.03)
Income	−0.10***	0.01	−0.06***	−0.03**
	(0.01)	(0.01)	(0.01)	(0.01)
Nonwhite	0.29***	−0.27***	−0.05	−0.06
	(0.05)	(0.04)	(0.05)	(0.04)
R^2	0.08	0.03	0.06	0.03
N	2953	2857	2894	2887

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes all particpants that passed the attention checks. Regression models predicting racial attitudes based on their treatment condition. In all instances, higher values indicate more anti-black racism.Model 1 (Date) uses the belief that it is acceptable for whites and blacks to date each other as the dependent variable. Model 2 (Dehumanization) uses the dehumanization scale as the dependent variable. Model 3 (Marry) uses respondents' reported comfort with a close relative marrying a black person as the dependent variable. Model 4 (Announcer opinions) uses a scale constructed from 4 items measuring opinions regarding the announcer as the dependent variable.

Table C.10: Impact of post/comments and racial resentment level on attitudes and opinions of all participants

	Model 1	Model 2	Model 3	Model 4
Constant	1.55***	1.64***	1.43***	1.34***
	(0.10)	(0.08)	(0.09)	(0.08)
Tweet only	−0.06	−0.13*	−0.13	−0.02
	(0.08)	(0.05)	(0.07)	(0.05)
Normative comments	−0.00	−0.05	0.02	−0.04
	(0.08)	(0.06)	(0.07)	(0.05)
Mixed comments	−0.09	−0.08	−0.09	−0.00
	(0.08)	(0.05)	(0.07)	(0.05)
Counter-normative comments	−0.02	0.05	−0.09	0.02
	(0.08)	(0.06)	(0.07)	(0.05)
High racial resentment	0.55***	0.91***	1.12***	0.43***
	(0.08)	(0.08)	(0.08)	(0.07)
Bachelor's	0.38***	−0.23***	0.31***	0.10**
	(0.05)	(0.04)	(0.04)	(0.04)
30 – 44 years old	0.22**	0.05	0.24***	−0.01
	(0.07)	(0.06)	(0.06)	(0.06)
45 – 49 years old	0.29***	−0.03	0.21***	−0.02
	(0.06)	(0.05)	(0.06)	(0.05)
60+ years old	0.32***	−0.07	0.20**	−0.02
	(0.07)	(0.06)	(0.06)	(0.05)
Male	0.18***	0.05	0.30***	0.20***
	(0.04)	(0.03)	(0.04)	(0.03)
Income	−0.09***	0.02*	−0.05***	−0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Nonwhite	0.34***	−0.19***	0.05	−0.03
	(0.05)	(0.04)	(0.04)	(0.03)
Tweet only × high racial resentment	−0.01	0.09	0.10	0.01
	(0.12)	(0.10)	(0.11)	(0.09)
Normative comments × high racial resentment	−0.05	0.11	−0.01	0.02
	(0.12)	(0.10)	(0.11)	(0.09)
Mixed comments × high racial resentment	0.03	0.07	0.02	0.03
	(0.12)	(0.10)	(0.11)	(0.10)
Counter-normative comments × high racial resentment	−0.02	−0.13	0.05	−0.05
	(0.12)	(0.11)	(0.11)	(0.09)
R^2	0.14	0.26	0.33	0.10
N	2919	2841	2871	2855

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Includes all particpants that passed the attention checks. Regression models predicting racial attitudes based on their treatment condition and level of racial resentment. In all instances, higher values indicate more anti-black racism. Model 1 (Date) uses the belief that it is acceptable for whites and blacks to date each other as the dependent variable. Model 2 (Dehumanization) uses the dehumanization scale as the dependent variable. Model 3 (Marry) uses respondents' reported comfort with a close relative marrying a black person as the dependent variable. Model 4 (Announcer opinions) uses a scale constructed from 4 items measuring opinions regarding the announcer as the dependent variable.

C-1.3 Survey instrument

1. Please select your gender.
    - Male
    - Female
    - Other
2. Are you of Hispanic, Latino, or Spanish origin?
    - Yes
    - No
3. Please select your year of birth.
    - 2010 ... 1905
4. Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what?
    - Republican
    - Democrat
    - Independent
    - Something else
5. When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?
    - Very conservative
    - Somewhat conservative
    - Slightly conservative
    - Neither liberal nor conservative; Moderate
    - Slightly liberal
    - Somewhat liberal
    - Very liberal
6. How often do you follow what's going on in government/public affairs?
    - Never
    - Sometimes
    - Most of the time
    - Always
7. Please indicate the highest level of education you have received.

- Less than high school
- High school graduate
- Some college, but no degree
- Associate's degree
- Bachelor's degree
- Post graduate degree (MA, MBA, MD, JD, PhD, etc.)

Pre-Treatment Racial Attitudes

8. The following questions will ask your opinions regarding race.

   Please indicate if you strongly agree, somewhat agree, neither agree nor disagree, disagree somewhat, or disagree strongly with each of the following statements.

   - I am fearful of people of other races.
   - White people in the US have certain advantages because of the color of their skin.
   - Racial problems in the US are rare, isolated situations.
   - I am angry that racism exists.
   - I think it is all right for blacks and whites to date each other.
   - Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
   - Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.
   - Over the past few years, blacks have gotten less than they deserve.
   - It's really a matter of some people just not trying hard enough: if blacks would only try harder they could be just as well off as whites.
   - Government officials usually pay less attention to a request or complaint from a black person than from a white person.
   - Most blacks who receive money from welfare programs could get along without it if they tried.

Psychopathic Traits

For each statement, select the choice that describes you best. There are no right or wrong answers; choose the answer that best describes you. Remember, your answers are completely confidential.

For each statement, select the bubble for the choice that describes you best. There are no right or wrong answers; just choose the answer that best describes you.

Remember, your answers are completely confidential.

9. Please indicate if each of the following statements is true, somewhat true, somewhat false, or false for you.

   - I am optimistic more often than not.
   - I have no strong desire to parachute out of an airplane.
   - I am well-equipped to deal with stress.

- I get scared easily.
- I'm a born leader.
- I have a hard time making things turn out the way I want.
- I have a knack for influencing people.
- I function well in new situations, even when unprepared.
- I don't think of myself as talented.
- I'm afraid of far fewer things than most people.
- I can get over things that would traumatize others.
- It worries me to go into an unfamiliar situation without knowing all the details.
- I can convince people to do what I want.
- I don't like to take the lead in groups.
- It's easy to embarrass me.
- I stay away from physical danger as much as I can.
- I don't stack up well against most others.
- I never worry about making a fool of myself with others.
- I'm not very good at influencing people.
- It is important that you pay attention, please select "Somewhat true".

10. Please indicate if each of the following statements is true, somewhat true, somewhat false, or false for you.

    - How other people feel is important to me.
    - I would enjoy being in a high-speed chase.
    - I don't mind if someone I dislike gets hurt.
    - I sympathize with others' problems.
    - I enjoy a good physical fight.
    - I return insults.
    - It doesn't bother me to see someone else in pain.
    - I enjoy pushing people around sometimes.
    - I taunt people just to stir things up.
    - I don't see any point in worrying if what I do hurts someone else.
    - I am sensitive to the feelings of others.
    - I don't have much sympathy for people.
    - For me, honesty really is the best policy.
    - I've injured people to see them in pain.
    - I sometimes insult people on purpose to get a reaction from them.
    - Things are more fun if a little danger is involved.
    - I don't care much if what I do hurts others.

- It's easy for me to relate to other people's emotions.
- It doesn't bother me when people around me are hurting.

Please indicate if each of the following statements is true, somewhat true, somewhat false, or false for you.

- I often act on immediate needs.
- I've often missed things I promised to attend.
- My impulsive decisions have caused problems with loved ones.
- I have missed work without bothering to call in.
- I jump into things without thinking.
- I've gotten in trouble because I missed too much school.
- I have good control over myself.
- I have taken money from someone's purse or wallet without asking.
- People often abuse my trust.
- I keep appointments I make.
- I often get bored quickly and lose interest.
- I have conned people to get money from them.
- I get in trouble for not considering the consequences of my actions.
- I have taken items from a store without paying for them.
- I have a hard time waiting patiently for things I want.
- I have lost a friend because of irresponsible things I've done.
- Others have told me they are concerned about my lack of self-control.
- I have robbed someone.
- I have had problems at work because I was irresponsible.
- I have stolen something out of a vehicle.

Social media habits

11. How often do you use some form of social media (Twitter, Facebook, TikTok, Reddit, Snapchat, Youtube etc.)

    - Several times a day
    - Once a day
    - A few times a week
    - A few times a month
    - Once a month
    - Once every few months
    - Never

12. Please indicate how often you use the following social media platforms:

- Facebook
- Pinterest
- Instagram
- LinkedIn
- Twitter
- Snapchat
- YouTube
- WhatsApp
- Reddit
- TikTok
- Nextdoor

Options for each platform: Often (at least daily), Sometimes (once or twice a week), Rarely (once or twice a month), Never

13. How often do you get your news from social media?

    - Often (at least daily)
    - Sometimes (once or twice a week)
    - Rarely (once or twice a month)
    - Never

14. Have you ever witnessed any of the following behaviors directed at a particular person online? (Not including something directed at you) [Check all that apply]

    - Someone being called offensive names
    - Someone being physically threatened
    - Someone being harassed for a sustained period
    - Someone being stalked
    - Efforts to purposefully embarrass someone
    - Someone being sexually harassed
    - None of the above

15. Which, if any, of the following have occurred to you, personally, ONLINE? [Check all that apply]

    - Been called offensive names
    - Been physically threatened
    - Been harassed for a sustained period
    - Been stalked
    - Had someone try to purposefully embarrass you
    - Been sexually harassed
    - None of the above

16. How often do you interact with others on social media (i.e. reply or comment to posts)

    - Often (at least daily)
    - Sometimes (once or twice a week)
    - Rarely (once or twice a month)
    - Never

17. How often do you engage in arguments on social media?

    - Often (at least daily)
    - Sometimes (once or twice a week)
    - Rarely (once or twice a month)
    - Never

18. Please indicate if you strongly agree, somewhat agree, neither agree nor disagree, disagree somewhat, or disagree strongly with each of the following statements.

    - I have posted shocking content because it is funny.
    - I like to troll people in the comments section of social media websites.
    - I enjoy griefing other players in multiplayer games.
    - The more beautiful and pure a thing is, the more satisfying it is to corrupt.
    - I enjoy making jokes at the expense of others.
    - I enjoy upsetting people I do not personally know on the internet.
    - Although some people think my posts and comments are offensive, I think they are funny.

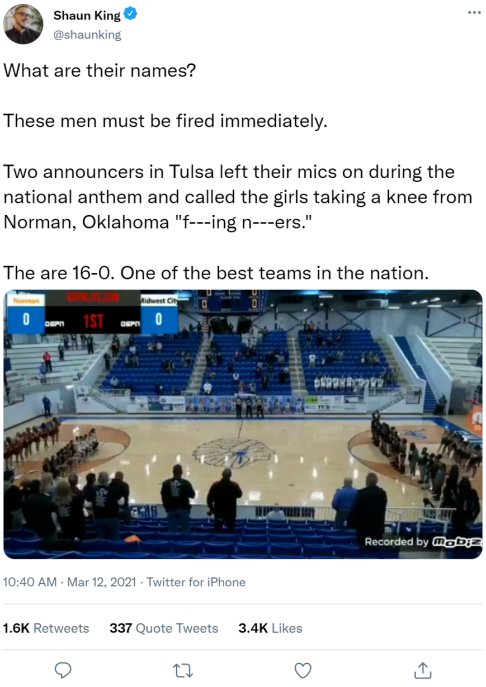
Interventions

19. Treatment groups

[The control group did NOT view the post nor its comments.]

[All treatment groups EXCEPT the control saw the following post:]

The next section will briefly describe an event in which racially insensitive language was used to describe high school students. This description will be in the form of a Tweet and may he reaction of Twitter users.



[The counter-normative comments:]

[The mixed comments:]

[The normative comments:]

Announcer attitudes

20. Recently, an announcer was caught on mic using a racial slur to describe high school basketball players who were kneeling during the national anthem.

    Please indicate if you strongly agree, somewhat agree, neither agree nor disagree, disagree somewhat, or disagree strongly with each of the following statements.

    - The announcer should be fired from his job.
    - The announcer was merely exercising his right to free speech.
    - Using racially insensitive language should always be punished.
    - Using racially insensitive language is often made into a bigger issue than it actually is.
    - Racial slurs should never be used.
    - No words should ever be banned.

Post-treatment racial attitudes

21. Please indicate if you strongly agree, somewhat agree, neither agree nor disagree, disagree somewhat, or disagree strongly with each of the following statements.

    Remember, your answers are completely confidential.

- I am fearful of people of other races.
- White people in the US have certain advantages because of the color of their skin.
- Racial problems in the US are rare, isolated situations.
- I am angry that racism exists.
- I think it is all right for blacks and whites to date each other.
- Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
- Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class.
- Over the past few years, blacks have gotten less than they deserve.
- It's really a matter of some people just not trying hard enough: if blacks would only try harder they could be just as well off as whites.
- Government officials usually pay less attention to a request or complaint from a black person than from a white person.
- Most blacks who receive money from welfare programs could get along without it if they tried.

22. Below are four different racial groups. Using the slider, indicate how comfortable you would be if a close family member were to marry an individual from the specified group (0 indicating you are not comfortable at all, 100 indicating you are completely comfortable).

    - White
        - 0 ... 100
    - Latino
        - 0 ... 100
    - Asian
        - 0 ... 100
    - Black
        - 0 ... 100

- How would you rate each group on a scale from 0 - 100, where 0 is not evolved or inhuman and 100 means fully evolved or human

People can vary in how human-like they seem. Some people seem highly evolved whereas others seem no different than lower animals. Using the image below, indicate using the sliders how evolved you consider the average member of each group to be:

  - White
    - 0 ... 100
  - Latino
    - 0 ... 100
  - Asian
    - 0 ... 100
  - Black
    - 0 ... 100

Post-treatment demographics

(a) Please indicate the highest level of education you have received.
  - Less than high school
  - High school graduate
  - Some college, but no degree
  - Associate's degree
  - Bachelor's degree
  - Post graduate degree (MA, MBA, MD, JD, PhD, etc.)

(b) Please indicate your annual income.
  - < $20,000
  - $20,000-$29,999
  - $30,000-$39,999
  - $40,000-$60,000
  - >$60,000