UNIVERSITY OF CALIFORNIA, IRVINE

Blinding Evaluations of Scientific Evidence Reveals and Reduces Partisan Biases

DISSERTATION

To be submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Psychological Science

by

Jared B. Celniker

DISSERTATION COMMITTEE:
Professor Peter H. Ditto, Chair
Associate Professor Paul K. Piff
Professor Linda J. Levine

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work was supported by the UCI Graduate Division Completion Fellowship and the National Science Foundation Graduate Research Fellowship Program.

I have been lucky receive mentorship from many amazing people throughout my graduate training. I am grateful to Dr. Chris Bauman for serving on my advancement committee and helping me narrow the scope of this dissertation. I am also appreciative of Dr. Linda Levine for taking her time and attention to serve on my advancement and defense committees, for which my work has already benefitted. Dr. Nathan Ballantyne has become an admired collaborator and friend, and I am thankful for the time he took to serve on my advancement committee. I am also indebted to Dean Gillian Hayes for her guidance and support, and I am excited to continue working with her in my new role at UCI. Yet there are three people who have shaped my interests and thinking more than an others over these past six years. To Dr. Paul Piff, thank you for your mentorship since my first days at UCI, my thinking and writing are sharper for having received your feedback throughout my training. To Dr. Azim Shariff, thank you for continually pushing me to think ambitiously and for supporting me to pursue those big ideas in a rigorous way. Working with you has made me more confident in my ideas and helped me recognize my ability to take on daunting challenges. To my committee chair, Dr. Peter Ditto, there's really no way to adequately express my gratitude for your mentorship and support. Thank you for taking a chance on me and for letting me roam the terrains of political and moral psychology to figure out my passions. This journey has been a lot of fun, and its inherent stresses have been minimized by knowing that you've always had my back.

I also would not have made it this far without the support of my friends. My Psych Science fam has been an essential well of support throughout these six years, especially my

cohort. Special shoutouts to Sean Goldy, Emma Grisham, John Michael Kelly, Becca Thompson, Nicky Jones, and my ex-house-husband Brett Mercier. Thank you all for helping me get through this ride and somehow making it enjoyable in the process. To all the friends I've made through AGS and other campus groups over these years, thank you for your service to your peers and for finding ways to band together even when everything seems stacked against us. Special shoutouts here to Maureen Purcell, Connor Strobel, Caitlin Suire, Melissa Dahlin, Yenda Prado, Kelli Malott, Janielle Vidal, and all the Pub Advisory Board members I served with! I'm also extremely lucky to have so many childhood friends still in my life. Thank you Nate West, AJ Sheiner, Nick Hanson, Alec Thimsen, Zane(zor) Waxman, Kyle Davis, Jacob Mitchell, Alex Stein, and Taylor Sexton for continuing to be a source of support and laughs.

My entire family has been instrumental in helping me get to this point. In particular, Ari and Seth Copeland, Gavin Fallen and Liz Landon, and Steve and Barb Fallen have been there for me through every step of this process. I'm incredibly grateful for all the family dinners and celebrations we've shared during my time in Irvine, and I never take your support for granted.

To my mom, Ilene Celniker, you have always been here for me, and I am only here because of all your hard work, love, and support. Thank you for always believing in me and for encouraging me to pursue my passions.

To Melody Moore, I am so grateful for the journey we are on together. You're my best friend and a better partner than I deserve. This adventure is ending, but I'm excited for all the ones we have ahead of us.

VITA

Jared B. Celniker

Education

2022	**Ph.D. in Psychological Science** Concentrations in *Social Psychology* and *Quantitative Methods* University of California, Irvine; School of Social Ecology
2019	**M.A. in Psychological Science** University of California, Irvine; School of Social Ecology
2014	**B.A. in Psychology**, *Summa Cum Laude* Academic Distinction in the Psychology and Honors Programs Chapman University; Crean School of Health and Life Sciences

Grants and Fellowships

2021 – 2022	**Graduate Completion Fellowship** ***($7,500)*** University of California, Irvine; Graduate Division
2017 – 2022	**NSF Graduate Research Fellowship** ***($138,000)*** National Science Foundation
2020	**Science in Action Fellowship** ***($500)*** University of California, Irvine; Graduate Division
2018	**Graduate Research Support Grant** ***($3,200)*** Charles Koch Foundation
2013	**Summer Undergraduate Research Fellowship** ***($5,000)*** Chapman University

Honors and Awards

2020	**Post-Baccalaureate Student Mentorship Award** ***($300)*** Department of Psychological Science
2019	**Social Glue Award** Department of Psychological Science
2019	**1st Place Oral Presentation** ***($500)*** AGS Research Symposium; University of California, Irvine

2016 **NSF Graduate Research Fellowship Honorable Mention**
National Science Foundation

2014 **Outstanding Senior**
Campus Leadership Awards; Chapman University

Publications

Bago, B., Aczel, B., Kekecs, Z., P., Kovacs, M., Nagy, T., … **Celniker, J.B.**,… Chartier, C. R. (in press). Situational factors shape moral judgments in the trolley dilemma in Eastern, Southern, and Western countries in a culturally diverse sample. *Nature Human Behavior*. https://doi.org/10.31234/osf.io/9uaqm

Mercier, B., **Celniker, J.B**., & Shariff, A.F. (2022*).* Overestimating explicit prejudice causes Democrats to believe disadvantaged groups are less electable. *Political Psychology.* https://doi.org/10.1111/pops.12820

Celniker, J.B., Ringel, M.M., Nelson, K., & Ditto, P.H. (2022*).* Correlates of "Coddling": Cognitive distortions predict safetyism-inspired beliefs, belief that words can harm, and trigger warning endorsement in college students. *Personality and Individual Differences.* https://doi.org/10.31234/osf.io/5g7nc

Kuchenbecker, S.Y., Pressman, S.D., **Celniker, J.B**., Grewen, K.M., Sumida, K.D., Naveen, J., Everett, B., Slavich, G.M. (2021). Oxytocin, cortisol, and cognitive control during acute and naturalistic stress. *Stress*, *24*(4), 370-383. https://doi.org/10.1080/10253890.2021.1876658

Ditto, P. H., Liu, B., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., **Celniker, J.B**., & Zinger, J. F. (2019). Partisan bias and its discontents. *Perspectives on Psychological Science, 14*(2), 304–316. https://doi.org/10.1177/1745691618817753

Ditto, P. H., Liu, B., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., **Celniker, J.B**., & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291. https://doi.org/10.1177/1745691617746796

Manuscripts Under Review

Celniker, J.B., Gregory, A., Koo, H., Piff, P. K., Ditto, P.H., & Shariff, A.F. (invited revision, *Journal of Experimental Psychology: General*). The moralization of effort. https://doi.org/10.31234/osf.io/nh9ax

Celniker, J.B., Rode, J.B**.,** Anderson, K.B., Ma, B., & Ditto, P.H. (invited revision, *Sex Roles*) College students' perceptions of ambiguous hook-ups involving alcohol intoxication.

Manuscripts in Preparation

Celniker, J.B., & Ditto, P.H. Blinding Evaluations of Science Reveals and Reduces Partisan Biases. (*Dissertation*)

Ballantyne, N., **Celniker, J.B.**, & Dunning, D. Do your own research.

Celniker, J.B., Grover, T., Tahk, R., & Ditto, P.H. Heterogeneity in moral profiles across self-reported political ideology.

Celniker, J.B., Ballantyne, N., **&** Ditto, P.H. Scared straight?: Shocking images are perceived as persuasive but do not change political attitudes.

Rodriguez, C.G., **Celniker, J.B,** & Ditto, P.H. The partisan lens in the beholding eye: Ideology and perceived victimhood in judgements of hate speech.

Manuscripts in Progress

Celniker, J.B., & Moore, M.M. Self-reported political ideology predicts ideal affect beyond ethnicity, income, and class.

Celniker, J.B., Spitz, S., & Ditto, P.H. Symmetries and asymmetries in conspiracism: Social conservatism and extreme partisanship predict conspiracy mentality.

Celniker, J.B., Stewart, B., & Benjamin, R. The influence of acute acetaminophen administration on political, moral, and prosocial judgments.

Presentations

Celniker, J.B. (2022). The moralization of effort. Flash talk at the Justice and Morality preconference at the Annual Convention of the Society of Personality and Social Psychology.

Ballantyne, N., **Celniker, J.B.**, & Dunning, D. (2021) Do your own research. Talk presented at The Philosophy of Epistemic Autonomy Conference.

Celniker, J.B. (2021). Approaches to moral and political psychology. Talk presented at the Baylor University Flourishing Brownbag.

Celniker, J.B., & Ditto, P.H. (2021). Methodological pre-commitments reveal and reduce political biases in evaluations of scientific information. Talk presented at the Society for Personality and Social Psychology Virtual Convention.

Mercier, B., **Celniker, J.B.,** & Shariff, A.F. (2021). Perceptions of prejudice and electability in the 2020 Democratic primary. Talk presented at the Society for Personality and Social Psychology Virtual Convention.

Celniker, J.B., Nelson, K., Ringel, M., & Ditto, P.H. (2020). Correlates of "Coddling" and cognitive distortions. Talk presented at the 2nd Biennial Heterodoxy in Psychology Conference.

Celniker, J.B. (2019). Studying partisans and informing policy: Two paths for social & behavioral scientists. Talk presented at the weekly meeting of the Science Policy Group @ UCI.

Celniker, J.B. (2018). The valorization of effort. Talk presented at the UCI Psychological Science Colloquium Series.

Conference Posters

Celniker, J.B, Nelson, K., Ringel, M., Ditto, P.H. (2020). Correlates of "Coddling" and cognitive distortions. Poster presented at the Annual Convention of the Society of Personality and Social Psychology.

Rodriguez, C.G., **Celniker, J.B,** Ditto, P.H. (2020). Hate speech is in the (left) eye of the beholder: Ideology, affinity, and perceived victimhood. Poster presented at the Annual Convention of the Association for Psychological Science.

Celniker, J.B, Rode, J.B**.,** Ditto, P.H. (2019). The influence of gender and intoxication on perceptions of sexual assault. Poster presented at the Annual Convention of the Society of Personality and Social Psychology.

Celniker, J.B, Shariff, A.F Piff, P. K., & Ditto, P.H. (2019). The moralization of effort: Policy implications of person perception. Poster presented at the Annual Convention of the Society of Personality and Social Psychology.

Celniker, J.B, Ditto, P.H. (2018). Not all cues are created equal: Partisan judgments influenced most by in-group opposition. Poster presented at the Annual Convention of the Society of Personality and Social Psychology.

Kuchenbecker, S. Y., Everett, B., **Celniker, J.B.**, Jonathan, N., Meija, L., Rojas, D., … & Loosle, D. (2017). Facial expressiveness and self-reported emotion: When the going gets tough. Poster presented at the Annual Meeting of Western Psychological Association.

Everett, B., **Celniker, J.B**., Jonathan, N., & Kuchenbecker, S. Y. (2017). Facial expressions and empathy: Acute stressful situations facilitate empathic response. Poster presented at the Annual Meeting of Western Psychological Association.

Kuchenbecker, S. Y., Everett, B., **Celniker, J.B**., Jonathan, N., Sumida, K., Slavich, G., & Pressman, S. (2016). When the going gets tough ~ Acute and Academic Stress: Natural oxytocin's association with negative emotion buffering, visual monitoring & cognitive

accuracy. Poster presented at the Annual Meeting of American Psychological Association.

Celniker, J.B., Everett, B., Slavich, G., Zimbardo, P., Colicino, C., Yoshiura, R., Gilbert, K., … Kuchenbecker, S. (2015). Silicone bracelets impact empathy, life satisfaction, past positive time perspective (ZTPI) & helping behavior. Poster presented at the Annual Convention of the Association for Psychological Science.

Jackson, L., Everett, B., **Celniker, J.B**., Lau, C., Robinson, M., Colicino, C., Kawai, S., … Kuchenbecker, S. Y. (2015). Cognitive interference in an emotionally evocative experience: Slowed RT and reduced discrimination accuracy. Poster presented at the Annual Meeting of Western Psychological Association.

Everett, B., **Celniker, J.B**., Jackson, L, Lau, C., Robinson, M., Colicino, C., Kawai, S., … Kuchenbecker, S. Y. (2015). Empathy, time perspective – past positive and future: Our well-being together. Poster presented at the Annual Meeting of Western Psychological Association.

Everett, B., **Celniker, J.B**., Colicino, C., Gilbert, K., Jacobsmeyer, A., Butterfield, C., Yoshiura, R., … Kuchenbecker, S.Y. (2014). Bystander helping behavior: Emotional empathic responsiveness associated with increased helping. Poster presented at the Annual Meeting of Western Psychological Association.

Celniker, J.B., Everett, B., Gilbert, K, Jacobsmeyer, A., Yoshiura, R., Butterfield, C. Silva, … & Kuchenbecker, S Y. (2014). Bystander helping and transituational factors: Gender, affective empathy and reminderbands. Poster presented at the Annual Meeting of Western Psychological Association.

Celniker, J.B., Everett, B., Gilbert, K., Colicino, C., Jacobsmeyer, A., Butterfield, C., Silva, H., … & Kuchenbecker, S.Y. (2013). Brief videos and "START everyday heroes" reminder wristbands facilitate well-being and helping behavior. Poster Presented at the Meeting of the International Positive Psychology Association.

Public Communication

Ballantyne, N., **Celniker, J.B., &** Ditto, P.H (2021) "Can Shocking Images Persuade Doubters of COVID's Dangers?" *Scientific American*. https://www.scientificamerican.com/article/can-shocking-images-persuade-doubters-of-covids-dangers/

Celniker, J.B. (2017). "Should You Donate to Disaster Relief?" *Psychology Today* and *The Decision Lab*. https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201708/should-you-donate-disaster-relief

Celniker, J.B. (2017). "Nudges: Social Engineering or Sensible Policy?" *Psychology Today* and *The Decision Lab.* https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201702/nudges-social-engineering-or-sensible-policy

Celniker, J.B. (2017). "Overcoming the Allure of Fake News." *Psychology Today* and *The Decision Lab.* https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201612/overcoming-the-allure-fake-news

Celniker, J.B. (2016). "Science Denial Isn't Only a Conservative Problem." *Psychology Today* and *The Decision Lab.* https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201610/science-denial-isn-t-only-conservative

Celniker, J.B. (2017). "To Be Right or Liked?" *Psychology Today* and *The Decision Lab.* https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201610/be-right-or-liked

Celniker, J.B. (2017). "Drone Policy: Reducing the Human Cost." *Psychology Today* and *The Decision Lab.* https://www.psychologytoday.com/us/blog/partisan-pitfalls-and-moral-misperceptions/201610/drone-policy-reducing-the-human-cost

Service and Leadership

Systemwide Service - University of California

2020	**Student Regent Designate Interview Committee**

Administrative and Advisory Committees - University of California, Irvine

2019 – 2021	**Student Health Insurance Advisory Committee**
2020	**Data Use Policy Task Force**
2020	**Semesters vs. Quarters Task Group**
2019 – 2020	**Coordinated Governance Group**

Academic Senate - University of California, Irvine

2021 – 2022	**Graduate Council**
2019 – 2022	**Council on Planning and Budget**
2017 – 2021	**Council on Faculty Welfare, Diversity, and Academic Freedom**

Associated Graduate Students - University of California, Irvine

2017 – 2022	**Graduate Student Representative, School of Social Ecology**
2019 – 2020	**Vice President of Internal Affairs**
2018 – 2019	**Pub Advisory Board Director**

School of Social Ecology - University of California, Irvine

2018 – 2021 **NSF Graduate Research Fellowship Review Committee**

2020 **Planning Group Member for Graduate Student Association**

2018 – 2020 **NSF Graduate Research Fellowship Student Q&A Panel**

Department of Psychological Science - University of California, Irvine

2018 – 2020 **Cohort Liaison Initiative Committee Representative**

2016 – 2018 **Graduate Student Panelist for Post-Baccalaureate Workshop**

Additional Service and Leadership

2018 – 2022 "**What Matters to Me and Why" Planning Committee Member**

2018 – 2019 **Science Policy Group @ UCI Executive Board**

2013 – 2014 **President, Gamma Beta Phi Honor Society (Chapman University)**

Mentorship

2018 – 2022 **Psychological Science Peer Mentor** to Desiree Chase, Julie Ann Kircher, Benjamin Kaveladze, & Madeline Snyder

2019 – 2021 **Undergraduate Student Mentor** to Brianna Ma and Veronica Chou (both won UCI UROP Funding Awards during mentorship)

2018 – 2020 **Post-Baccalaureate Student Mentor** to Karli Nelson

Additional Experiences and Funding

2018 – 2021 **AGS Travel and Conference Grants *(totaling $1050)***

2020 **Summer Research Appointment, UCI Graduate Division**

2015 **Research Internship, The Center for Responsive Politics**

2008 - 2018 **Cofounder & Trustee, A World of Good, a 501(c)(3) organization**

ABSTRACT OF THE DISSERTATION

Blinding Evaluations of Scientific Evidence Reveals and Reduces Partisan Biases

By Jared B. Celniker

Doctor of Philosophy in Psychological Science

University of California, Irvine, 2022

Professor Peter H. Ditto, Chair

Do political partisans evaluate new information in a biased way? Despite decades of research, this question has been difficult for psychologists to resolve. Proponents of rationalist accounts claim that ostensible biases can be solely explained by impartial, accuracy-motivated processes; in contrast, proponents of motivated accounts contend that partisans' evaluations are often influenced by biased, directionally-motivated processes. Embracing the logic of *blinding* that underlies many scientific practices, I designed and deployed a novel experimental paradigm across four preregistered experiments (N = 4010) to critically test these two accounts. Participants were randomly assigned to evaluate the methodological quality of scientific evidence either *before* they knew its results (blinded evaluations) or *after* they knew its results (unblinded evaluations). The critical assumption underlying this design was that blinded participants provided purely impartial, accuracy-motivated evaluations. Unblinded participants, on the other hand, may or may not have been biased by their prior political beliefs when evaluating the new information. Indeed, in every study, unblinded participants were unduly influenced by their prior beliefs compared to their attitudinally-similar counterparts who made blinded evaluations. Partisans' feelings and expectations produced independent, yet highly intertwined, biasing effects on their evaluations. These biases were most evident in unblinded

participants' denigration of politically-unfriendly information, rather than their veneration of politically-friendly information. Additionally, I found that blinding partisans' quality evaluations influenced how credible they found politically-unfriendly information and how they updated their beliefs in response to that information. I discuss how these findings can be integrated into existing theoretical models—including Bayesian models of belief updating—to provide more accurate descriptions of political cognition. Ultimately, these results disconfirmed strong rationalist accounts of partisan evaluations and supported the existence of partisan biases.

INTRODUCTION

"It ain't what you don't know that gets you in trouble. It's what you know for sure that just ain't so."

Imagine a friend who believes the quotation above is originally attributable to Mark Twain. The tidbit may have been relayed to him through a trusted family member or coworker, or perhaps he came across it through a movie or social media post. The quote resonated with him, and he has recited it often throughout the intervening years, noting the wisdom in Twain's wit in such recitations. Yet after stumbling across an expert resource on the quotation (Seybold, 2016), you have been convinced that the attribution is apocryphal. You share this information with your friend, but he won't budge in his belief that the quote is Twain's. He doesn't find the archival evidence very rigorous, especially since it relies on what The New York Times *couldn't* find. Despite your evidence, he maintains his belief in the quote's origins, and you hear him misinform others about Twain's alleged aphorism in subsequent conversations. After a spat or two on the topic, you frustratedly let it go, wincing whenever he brings it up but knowing your efforts are better spent elsewhere.

Should we consider your friend's resistance to new information to be biased? Although the answer seems obvious, it is a difficult question to parse. Psychologists have long struggled to conceptually and empirically differentiate between judgments that are biased and ones that are merely flawed. This has been called the *observational equivalence problem*: what often looks like a bias tends to be indistinguishable from a judgment one would make via impartial, albeit imperfect, judgment (Tetlock & Levi, 1982; Ditto, 2009; Druckman & McGrath, 2019). Your friend's questioning of archival research methods may have merit, or he may have good reasons for trusting the original source more than the university at which the Center for Mark Twain

Studies is housed; it is also possible that, having recited the misinformation over years, your friend does not want to believe that he could have uncritically spread such misconceptions.

The case of the apocryphal quote is fairly low-stakes, yet the difficulty of distinguishing biased from unbiased reasoning is prevalent across domains of human judgment (Ditto, 2009; Erdeyli, 1974; Gerber & Green, 1999; Miller & Ross, 1975; Tetlock & Levi, 1982). Two domains in which these distinctions are consequential are science (Lord et al., 1979; Munro & Munro, 2014; Scurich & Shniderman, 2014; Washburn & Skitka, 2017) and politics (Baron & Jost, 2019; Ditto et al., 2019a; Druckman & McGrath, 2019; Kahan, 2016; Leeper & Slothus, 2014; Tappin, et al., 2020c). Whether people are biased or unbiased evaluators of political information, and politically-relevant scientific evidence, may have practical implications for how policymakers, social media platforms, and everyday citizens seek to convey such information to others. In an era of historic polarization (Druckman, 2017; Finkel et al., 2020), understanding how people process information and change their beliefs—particularly after encountering information that contradicts what they previously believed—is critical for cultivating social consensus to address collective problems.

Defining Bias

Historically, researchers of social cognition have struggled to distinguish between judgments made through rational, unbiased information processing and those made through irrational, biased information processing. Debates over the rationality or irrationality of judgment has taken many forms (for a historical overview, see Ditto, 2009), but it can be distilled into a debate over the extent to which desires and expectations guide how people evaluate information. For instance, recognizing threatening stimuli more quickly than nonthreatening stimuli, or attributing successes to oneself and failures to external influences, can be thought of as

demonstrating a motivational bias that drives people to process information in accordance with their desires. Yet these results can be explained similarly plausibly as violations of expectations; people do not expect to see threatening stimuli or fail at a task, and people may process unexpected information and expected information differently regardless of their underlying motivations (Erdeyli, 1974; Miller & Ross, 1975; Tetlock & Levi, 1982; Gerber & Green, 1999; Baron & Jost, 2019). The seminal work on cognitive heuristics separately demonstrated that it is unnecessary to invoke motivational processes to explain many predictable errors in judgment (Kahneman, 2003). Thus, while early motivational theorists argued that desires and preferences bias judgments, alternative accounts of "unmotivated" information processing could not be disproven.

In response to these empirical and conceptual criticisms, motivational theorists developed integrative frameworks that conceptualized motivations as explicitly influencing cognitive processing (Kunda, 1990; Ditto & Lopez, 1992; Kruglanski, et al., 2009; Ditto, 2009). Importantly, the latest generation of motivational theories acknowledged that both "hot" (affective) and "cold" (cognitive) psychological processes influence people's judgments (Kunda, 1990; Ditto & Lopez, 1992; Ditto, 2009). The key insight from these updated motivational theories was that all reasoning is motivated—that is, knowing the relative contributions of "hot" and "cold" systems (to the extent they are dissociable) is only useful insofar as it can help researchers detect the *types* of motivations that are driving a particular judgment. Many of these updated motivational theories focused on differentiating two broad classes of motivations that people hold, typically simultaneously: accuracy and directional motivations. Accuracy motivations are characterized by the drive to obtain an *accurate* understanding of the world, while directional motivations reflect the drive to maintain or defend a *particular* understanding

of the world (Kunda, 1990; Leeper & Slothuus, 2014). Directional motivations can alter how information is processed by, for example, influencing how much attention one pays to belief-inconsistent, relative to belief-consistent, information (Ditto & Lopez, 1992; Ditto, et al., 1998). This conceptualization of motivations has proven useful for researchers interested in defining and measuring political biases. A partisan bias has been defined as the motivation to arrive at politically-friendly (i.e., belief-consistent) conclusions, in contrast to the motivation to interpret information as accurately as possible (Bolsen et al., 2014; Druckman & McGrath, 2019; Leeper & Slothus, 2014; Kahan, 2016; Tappin, et al., 2020c).

Limitations of Past Experimental Research on Partisan Bias

Nevertheless, studies of partisan bias have continued to suffer from the problem of observational equivalence, making it difficult to reinterpret past research on partisan bias through the lens of these theoretical developments. Our recent meta-analysis reanalyzed the results of 51 paradigmatic experiments on the topic from political science and political psychology (Ditto et al., 2019a). To be included in the meta-analysis, studies were required to present tightly matched information to participants (either between- or within-subjects) that manipulated the political friendliness of the information presented. Research that met these criteria tended to fall into two classes of experimental designs: "party cue" designs that manipulate the source of presented information across participants, and "outcome switching" designs that manipulate the content of presented information across participants. Although these experimental designs are common, many of these studies cannot definitively exclude rational, accuracy-motived counterexplanations of purportedly biased, directionally-motivated effects (Baron & Jost, 2019; Tappin et al, 2020c).

For example, in a prototypical "party cue" experiment, participants are presented with the same proposed policy (for example, a welfare policy), but the political party supporting the legislation is manipulated across conditions (e.g., Cohen, 2003). Participants then complete ratings of how good or bad they think the policy is and how much they support or oppose it. The logic of this design is that, if partisans support the exact same policy more when it is endorsed by their preferred party (compared to the opposing party or no endorsement), then their ratings belie a directional motivation to support their party rather than an accuracy motivation to interpret the policy evenhandedly. While many researchers have argued that effects observed in Party Cue experiments demonstrate a bias, others note that the same pattern of behavior can be explained without invoking bias. For example, partisans may contextualize complicated information through the strong, informative signal of their partisan affiliations. If people align themselves with a particular political party because they believe that party is usually correct, then it can be reasonable and efficient to use a party cue as a guide for assessing support for a policy (Bullock, 2009).

Similarly, in a prototypical "outcome switching" design, all participants are presented with information about a political topic (e.g., a summary of a research article testing the effectiveness of capital punishment), and only the content of the information is manipulated across conditions (e.g., results suggest the policy is/is not effective; Lord, Ross & Lepper, 1979). Participants then rate the quality of the study, whether they have changed their attitude on the topic after reading about the study, and other dependent measures. The logic here is that, if partisans think a study is higher quality or stronger evidence in support of a conclusion when the results are politically-friendly (compared to politically-unfriendly results), then they are demonstrating a directional motivation to align their evaluations of the study with their prior

views rather than update their beliefs in an accuracy-motivated manner. These matched-information designs have been the gold-standard methodology for assessing biases, political or otherwise, following precedents set by studies on heuristics in judgment and decision-making research (Kahneman, 2003; Ditto et al., 2019b). Yet researchers have noted the ostensible biases observed in outcome switching experiments may be attributed to accuracy-motivated partisans forming and changing their beliefs through the lens of their prior attitudes. Indeed, partisans' prior attitudes were likely developed through exposure to trusted information and sources. If people come to their beliefs about a political issue by engaging with what they think is the best evidence on the subject, it can be reasonable to dismiss the results of a contradictory study as a statistical fluke or as coming from an untrustworthy source (Baron & Jost, 2019; Tappin, et al., 2020c).

Thus, despite the clear logic of these two experimental paradigms, the observational equivalence problem has made it difficult to derive clear insights about the prevalence of partisan biases from these studies. One cannot determine whether a partisan is accuracy-motivated (seeking the truth regardless of political-friendliness) or directionally-motivated (seeking a politically-friendly conclusion regardless of whether it is true) by solely measuring their prior attitudes and their post-information beliefs (Bullock, 2009; Ditto, 2009; Leeper & Slothus, 2014; Baron & Jost, 2019; Tappin et al., 2020c). Accuracy-motivated accounts of partisan judgments, and rationalist accounts of human judgment more broadly (Jones & Love, 2011), highlight that it is not inherently biased for people to use their prior beliefs as a filter through which to understand the world. Indeed, many have argued that evaluating new information in reference to one's prior beliefs is ideal for, if not integral to, learning and accurate belief-updating (Baron & Jost, 2019; Druckman & McGrath, 2019; Gerber & Green, 1999; Lord, Ross & Lepper, 1979).

While people will sometimes reach inaccurate conclusions by filtering and interpreting new information through the lens of their prior beliefs, engaging in such a process may actually lead to more accuracy in the long run.

Motivations and Bias in a Bayesian Framework

Although the observational equivalence problem makes it harder to clearly identify directional biases, not all types of judgments are equally susceptible to accuracy-motivated counterexplanations. Evaluations of the quality of new information, as is often assessed in outcome switching designs, are not as open to rationalist counterexplanations as are impressions of the credibility of that information or measures of belief updating (Ditto et al., 2019a, 2019b). To better understand the often fine distinctions between *quality evaluations*, *credibility impressions*, and *belief updating*, it may be useful to contextualize these concepts in a Bayesian belief updating theoretical framework.

Bayesian belief updating is a theoretical process through which people incorporate new information into their prior beliefs to arrive at a new, or updated, beliefs (Druckman & McGrath, 2019; Jones & Love, 2011). People start off with belief, which are called prior beliefs (often referred to simply as priors). People hold many prior beliefs with varying amounts of confidence. When people encounter new information, they must interpret it. I refer to this second process as evaluation. Evaluation entails an assessment of the quality or validity of the new information (e.g., if a scientific study seems methodologically sound). People also form impressions of the credibility of the new information (e.g., if the information comes from a trustworthy source), which may occur before or after evaluating the quality of that information. Finally, people assimilate (or do not assimilate) the new information into their prior beliefs, the process of belief updating, through which they arrive at a new, updated belief.

The extent to which an updated belief differs from a prior belief is a function of both the perceptions of the new information and the level of confidence with which a person held their prior belief. If a person was not confident in their prior belief, then it may not take particularly high-quality information for them to arrive at an updated belief in the direction of the new information. Yet if a person was extremely confident in their prior belief, then only the highest-quality information that contradicts their prior belief may suffice for driving the reasoner to arrive at an updated belief that meaningfully deviates from their prior belief. Critically, a Bayesian belief-updating process can be driven by a reasoner holding accuracy and/or directional motivations (Druckman & McGrath, 2019). Although Bayesian models are often referenced as rationalist or normative models of reasoning (Baron & Jost, 2019), even by critics (Jones & Love, 2011; Uhlmann, 2011), a Bayesian belief updating process can result in directional biases.

In theory, Bayesian belief updating can go awry at the stage of quality evaluations, credibility impressions, belief updating, or any combination thereof. It is extremely difficult to tease apart whether a directional bias has emerged in the credibility impression or belief updating stages of reasoning. Expectations (accuracy motivations) and preferences (directional motivations) are often intertwined and have similar effects on judgments. To provide evidence of a directional bias in one's credibility impressions or belief updating, researchers must show that a failure to believe new information is unequivocally driven by a person's preferences (see Thaler, 2019 for recent methodological advances in this area). Although the same standards apply to detecting biases in evaluations, it is much more straightforward to detect directional biases in the evaluation stage (Ditto et al., 2019a, 2019b). This is because an accuracy-motivated reasoner would not use the congeniality of new information with their prior beliefs to assess the quality of that information. If a person is motivated to come to an accurate understanding of the

world, the political-friendliness of new information is irrelevant to assessing its quality (Kelly, 2008, Ballantyne, 2019, Carter & McKenna, 2020). While prior beliefs can (and, in fact, almost inevitably will) affect credibility impressions and belief updating in a purely accuracy-motivated person (Lord, Ross & Lepper, 1979; Druckman & McGrath, 2019), that same accuracy-motivated individual would not be influenced by their priors in evaluating the quality of the information in question. On the other hand, if a person is motivated to come to a *particular* understanding of the world, then the political-friendliness (i.e., belief-consistency) of the new information becomes relevant to understanding how valid it may be. In other words, the prior beliefs of an accuracy-motivated partisan would be orthogonal to their quality evaluations, but a partisan influenced by directional motivations would produce evaluations that are associated with their prior beliefs to some degree.

Evaluations of Scientific Methodology are Ideal Outcomes for Measuring Partisan Bias

In a critique of our meta-analysis, some researchers contended that quality evaluations, including evaluations of scientific methodology, can be influenced by one's prior beliefs in a way that is consistent with accuracy-motivated, normative reasoning (Baron & Jost, 2019). The argument here is that, if the results of new information conflict with one's prior beliefs, then it can be rational and accuracy-motivated to denigrate the quality of such information when one is confident that their prior beliefs are correct. In Bayesian terms, if a person is 100% certain in their prior belief, then new information that contradicts that belief can be evaluated as being of lower quality than otherwise identical information that affirms one's prior belief. If this argument is sound, then using quality evaluations as the focal outcome variable to measure partisan bias may be as challenging as using credibility impressions or belief updating outcomes.

However, there are good reasons, beyond the aphorism that opened this dissertation, to be skeptical of this argument. Even if one grants that the argument holds in the limiting case, where the true state of the world is objectively known, the argument immediately falls apart as soon as there is any doubt about the true state of reality—as is almost always the case. When there is any uncertainty about the true state of the world, denigrating the quality of new information that does not fit with one's prior beliefs can create a closed circuit that inhibits one from developing accurate beliefs (Ditto et al., 2019b). In other words, the political-friendliness of new information does not reflect the quality of that information.

The disjunction between the political-friendliness and quality of new information is evident when considering evaluations of the methodological quality of scientific evidence. While researchers have sometimes used outcome measures that are too general and may reflect an evaluation of methodological quality or an impression of its credibility (i.e., "how convincing" one found a study, Lord et al., 1979), it is easier to discern the type of judgment being made when one deploys more specific items to measure these outcomes (e.g., Munro & Munro, 2014; for similar arguments for discussions of argument quality, see Areni & Lutz, 1988). For example, it is difficult to create rational, unbiased explanations for evaluating a specific sample size as being less adequate when a study yields politically-unfriendly, compared to politically-friendly, results. Similarly, whether a study produces evidence for or against the deterrent effect of capital punishment should not influence how well one thinks that the murder rate (or some other outcome variable) captures the prevalence of violent crime. Such ideological epistemology, shifting standards of evidence in relation to one's ideological commitments, is not normatively defensible nor compatible with accuracy-motivated accounts of judgment (Kelly, 2008; Ballantyne, 2019; Clark & Winegard, 2020; Carter & McKenna, 2020). In sum, evaluating the

quality of new information through the lens of one's prior beliefs has been explicitly recognized as a directional bias in the evaluation of new information (e.g., prior attitude effect, Druckman & McGrath, 2019).

The Logic of Blinding

The recognition that one's prior beliefs can bias their evaluations underpins many strategies, including common scientific practices, to prevent bias. *Blinding* strategies, which restrict an evaluator's access to information which has been deemed irrelevant to the evaluative outcome, have been shown to reduce biases in contexts as diverse as orchestral job interviews (Goldin & Rouse, 2000) to clinical trials (Schultz et al., 1995). In the absence of blinding, evaluators can be "contaminated" (Wilson & Brekke, 1994) by normatively irrelevant information, like the gender of a job applicant. Contamination can result in the reconstruction of evaluative criteria to favor preferred evaluative outcomes, like favoring experience (or education) in job hiring to make one's evaluative criteria match the strengths of a preferred job candidate (Uhlmann & Cohen, 2005). In science, the logic of blinding continues to guide reforms to improve quality and replicability (e.g., registered reports, Chambers et al., 2015; Munafò et al., 2017; Nosek & Lakens, 2014). One's evaluations of the quality of a study cannot be biased by its results without knowing them (Ebersole, 2019), so registered reports help reduce publication biases by publishing methodologically rigorous research regardless of whether its results are statistically significant. Indeed, scientists have institutionalized the logic of blinding at many levels of the peer-review process to protect evaluations of methodology from reviewers' potential biases.

The blinding paradigm

Blinding has been applied in a variety of domains to prevent bias. Surprisingly, blinding has not been used to help resolve debates over the existence of partisan bias. Using blinding as a feature of experimental designs can help redress limitations of past research in this area. In between-subjects experimental designs, partisans can be randomly assigned to evaluate specific aspects of a study's methodological quality either *before* the results of the study are presented (blinded) or *after* the results are known (unblinded). This design mirrors those of outcome switching experiments, but participants in a blinding paradigm evaluate identical methods with the results either presented or withheld. The critical assumption underlying this design is that participants assigned to commit blinded study quality evaluations are accuracy-motivated when making those evaluations (Bastardi, Uhlmann & Ross, 2011). Without knowing a study's results, any directional goals that partisans may hold cannot bias those initial evaluations (Ebersole, 2019). In contrast, participants assigned to provide unblinded evaluations of study quality may act on both their accuracy and/or directional motivations in evaluating the study. Establishing accuracy-motivated baselines of partisans' blinded evaluations can be used as the normative comparison against partisans' unblinded, and potentially biased, evaluations.

The role of prior beliefs

A notable limitation of past work on partisan bias has been the lack of measurement of prior beliefs to account for their influence on evaluations (Druckman & McGrath, 2019; Tappin et al., 2020c). Throughout the remainder of this dissertation, I will use the term *prior beliefs* to collectively refer to the broad range of beliefs, feelings, preferences, and expectations that may influence partisans' evaluations; when referring to specific beliefs, I will use more specific terms (e.g., prior support, prior efficacy beliefs, partisan feelings). To clearly show a partisan bias, it is necessary to show what prior beliefs may be exerting directional influence on partisans'

evaluations. To maximize the ability of the blinding paradigm to shed light on outstanding questions about partisan bias, it is critical to also measure and model partisans' prior beliefs. Measuring prior beliefs, and modeling their influence across blinded and unblinded conditions, can clarify the extent to which knowing the results of a politically-relevant study (i.e., how politically-friendly the results are) biases unblinded evaluations. While blinded partisans' prior beliefs should be orthogonal to their quality evaluations, unblinded partisans' quality evaluations may or may not be directionally influenced by their prior beliefs.

Measuring and modelling prior beliefs can also help address three other limitations of past research on partisan bias. First, rather than assuming co-partisans share the same beliefs about a particular topic, measuring specific prior beliefs can allow researchers to compare partisans who truly have similar attitudes about a topic. Comparing individuals with similar prior beliefs can "[hold] tastes constant" (Gerber & Green, 1999, pg. 206) more than in past research and thus enable more precise estimation of partisan biases (Tappin et al., 2020c).

Critically, these benefits to bias estimation hold even if there are measurement errors in capturing partisans' prior beliefs. For example, it is plausible that partisans' differ in their preferences for certain types of empirical evidence (e.g., favoring naturalistic correlations to laboratory experiments) or with their preferences for different kinds of evidence more broadly (e.g., empirical vs. anecdotal evidence; Kubin et al., 2021), and such epistemic preferences may be captured in measures of their prior beliefs. It is typically assumed that partisans agree upon the quality of different types of evidence when political motivations are not in play, but partisans may differ in their epistemic preferences. There are no clear a priori predictions about how partisans may differ in their epistemic preferences, and the few studies that have examined this assumption have found mixed evidence (Munro & Munro, 2014; Scurich & Shniderman, 2014).

Nevertheless, blinding the evaluations of participants with *different* prior beliefs provides a test of whether they differ in their epistemic preferences, and comparing the blinded and unblinded evaluations of participants with *similar* prior beliefs can still yield estimations of bias even if the belief measure captures differences in epistemic preferences. By measuring prior beliefs prior to random assignment, there should be equal measurement error across conditions, so comparing the evaluations of blinded and unblinded co-partisans can still yield evidence of directional bias. Thus, measuring and modeling the influence of partisans' prior beliefs on their quality evaluations allows researchers to estimate partisan bias even when there is evidence of measurement error.

Second, measuring participants' prior beliefs also enables researchers to test the types of beliefs and individual differences that may be driving partisan biases. Some theorists of political cognition have argued that partisans judgments are biased by their feelings and social concerns (Kahan, 2016), but others contend that differences in partisan judgment are more readily explained by differences in specific expectations and cognitive capacities (Tappin et al., 2020a, 2020b). It is important to reiterate that it is not necessary to dissociate the contributions of affect and cognition to measure a bias: feelings and expectations can each exert directional influences on partisan evaluations (Druckman & McGrath, 2019). Nonetheless, identifying the psychological drivers of biases can help researchers more readily target personality and social factors that are relevant to partisan biases. If partisan biases are more cognitively-driven, then prior beliefs about the efficacy of a policy (e.g., the effectiveness of capital punishment for reducing violent crime) should have a stronger influence on partisans' unblinded evaluations than their prior support for that policy (e.g., how much they support or oppose capital punishment). Alternatively, if partisan biases are more affectively-driven, then measures like

partisans' prior support should exert greater biasing influences than their prior efficacy beliefs. Understanding the extent to which partisans' evaluations are cognitively- or affectively-driven will clarify the types of individual differences, from cognitive dispositions like analytic thinking (Batailler et al., 2022; Kahan et al., 2017; Pennycook & Rand, 2019; Tappin et al., 2020a) and intellectual humility (Bowes et al., 2022) to affective dispositions like one's moral conviction (Skitka, et al., 2005; Skitka et al., 2012; Skitka & Wisneski, 2011) and attachment to a political party (Kahan, 2016), that may mitigate (or exacerbate) partisan biases. The blinding paradigm can afford strong tests of competing theoretical accounts by allowing researchers to assess the influence that various prior beliefs and individual dispositions have on partisans' quality evaluations.

Third, measuring partisans' prior beliefs allows researchers to model the influence that such beliefs have on the entire information processing stream—and how such processing may differ as a function of knowing the results when one makes their quality evaluations. Despite the abundance of theoretical models of belief updating (Jones & Love, 2011; Tappin et al., 2020c), few studies attempt to descriptively map the belief updating processes. Estimating how partisans' prior beliefs influence their quality evaluations, and how their prior beliefs and quality evaluations subsequently influence their credibility impressions and belief updating, can provide more descriptive information about how partisans process information than is typically assessed. Such information is essential for understanding the viability of blinding strategies for fostering convergent belief updating and, ultimately, consensus after evaluating shared information.

Overview of the current research

In this dissertation, I sought to measure the extent to which partisans' prior beliefs influence their judgments of politically-relevant scientific evidence. I conducted four

preregistered experiments using the blinding paradigm to address three research questions about partisan bias.

Research Question 1: Do partisan biases exist?

The central focus of these studies was to provide dispositive evidence against purely rationalist, accuracy-motivated accounts of partisan judgment. By randomly assigning participants to commit blinded or unblinded evaluations—assessing the methodological quality of a study either before or after they know its results—and measuring how their prior beliefs influence their evaluations, I tested whether unblinded participants' quality evaluations were influenced significantly more by their prior beliefs than blinded participants. I hypothesized that partisans' prior beliefs would significantly influence unblinded participants' evaluations but not blinded participants' evaluations, indicating a directional influence of partisans' prior beliefs on their evaluations. In each study, I also explored whether other measures of partisan motivations (e.g., political orientation) biased unblinded participants' evaluations.

Research Question 2: If partisan biases exist, what drives them?

In each study, I collected at least two measures of participants' prior beliefs, one conceptualized as being more reflective of participants' feelings (prior support in Studies 1a-2, partisan feelings in Study 3) and one conceptualized as being more reflective of participants' expectations (prior efficacy beliefs in Studies 1a-2, prior beliefs about partisan bias in Study 3). This allowed me to test whether those specific beliefs exerted directional influences on participants' quality evaluations. In each study, I hypothesized that both measures of prior beliefs would bias partisans' quality evaluations. Additionally, I explored the relative influence that each prior belief measure had on quality evaluations when controlling for the influence of the other. In Study 2, I tested whether differences in positive and negative affect could account for differences

in evaluations across the blinded and unblinded conditions. In Studies 2 and 3, I also examined whether various individual difference measures (e.g., moral conviction, analytic thinking) exacerbated or mitigated any observed partisan biases.

Research Question 3: How does blinding partisans' quality evaluations influence their subsequent credibility impressions and belief-updating?

In addition to developing more specific measures of methodological quality than have been used in prior research, I created measures of credibility impressions and used measures of belief updating (i.e., attitude change) to assess how blinding quality evaluations influenced the information processing stream. I did not have specific hypotheses regarding this research question, but the analyses I conducted were designed to test whether blinding participants' quality evaluations led to them form more positive credibility impressions and update their beliefs more than unblinded participants.

The preregistrations for these studies are available on my OSF page (https://osf.io/xhjk7/?view_only=de21e005c2fc41d9a70536b1516a5330). To highlight the analyses most relevant to these main research questions, many of the preregistered analyses are presented in the appendices. I deviated from some of my preregistered analyses (particularly in the analyses related to Research Question 3) to present more theoretically relevant analyses than originally planned. In all cases, the results of the preregistered analyses do not substantively alter the conclusions drawn from the analyses and results presented in the main text.

STUDIES 1A AND 1B

After pilot testing summaries of scientific evidence and measures of study quality and credibility impressions in student samples, I conducted Studies 1a and 1b as initial tests of whether blinding reduces directional biases in evaluations of scientific evidence. In Study 1a, participants were asked to read materials, adapted from studies of biased assimilation (Liu, 2014; Lord, Ross & Lepper, 1979), summarizing an ostensible study about capital punishment reducing state-levels murder rates. In Study 1b, participants were asked to respond to materials, based on an actual experiment (Bellet et al., 2020), summarizing a study about how trigger warnings do not effectively reduce anxiety preceding exposure to distressing content.

Notably, unlike many outcome switching experiments that counterbalance the presentation of politically-friendly ad politically-unfriendly results, I used results that were consistently more politically-unfriendly to liberals in these studies. There were two main reasons for this design choice. First, using a single set of materials substantially decreased the number of participants I needed to recruit for adequately-powered analyses, which was a primary consideration in these initial studies. Second, online participant pools tend to skew liberal (Clifford et al., 2015; Levay et al., 2016), and I expected politically-unfriendly information to yield larger directional biases. Although this design choice inhibited me from conducting strong tests of (a)symmetries in partisan bias, it allowed me to more powerfully test whether liberals display partisan biases, a claim that has been contested in recent years (Baron & Jost, 2019).

Both Study 1a and 1b tested the same focal hypotheses. In relation to Research Question 1, I hypothesized that participants' prior beliefs would influence unblinded evaluations of study quality but not blinded evaluations of quality—revealing directional influences on participants' unblinded evaluations. In relation to Research Question 2, I hypothesized that both participants'

prior support and their prior efficacy beliefs would bias their unblinded quality evaluations, and I examined the unique variance explained by each measure when accounting for the influence of the other. In relation to Research Question 3, I examined whether the blinding manipulation influenced how participants formed impressions of the credibility of the study (e.g., how believable they found results were) and how they updated their beliefs after considering the presented evidence.

Method

Participants

A power analysis indicated that, to have 0.80 power to detect small ($f^2 = 0.03$) two-way interactions, I would need to recruit at least 434 participants per study. Anticipating that some participants would not pass our preregistered inclusion criteria (three English comprehension questions, a manipulation check, and taking more than three minutes to complete the survey), I aimed to recruit 500 participants per study through Prolific.

Ultimately, 501 and 500 participants were recruited for Study 1a and Study 1b, respectively, in January of 2021. In Study 1a, the participants ranged in age from 18-75 years ($M = 32.66$, $SD = 12.82$), ranged in yearly household income from less than $5,000 to over $175,000 (*Mdn* = $50,000 - $59,999) and were mainly White (63%), female (57%), and college-educated (36%). In Study 1b, the participants ranged in age from 18-74 years ($M = 32.35$, $SD = 12.09$), ranged in yearly household income from less than $5,000 to over $175,000 (*Mdn* = $50,000 - $59,999) and were mainly White (64%), male (52%), and college-educated (35%).

Procedure and Measures

After consenting, completing a captcha, and responding to three English comprehension questions, participants in each study were asked to complete measures of policy support, beliefs

about the efficacy of the policy (referred to henceforth as efficacy beliefs), and moral conviction. To measure policy support, participants were asked, "To what extent do you support or oppose [capital punishment in the United States/the use of trigger warnings]?" and responded on a 7-pt scale (1 = *Strongly oppose,* 4 = *Neither oppose nor support,* 7 = *Strongly support*). Participants were then asked, "To what extent do you believe that [capital punishment is an effective deterrent of violent crime/trigger warnings are effective at reducing anxiety for people exposed to distressing content]?" and responded on a 7-pt. scale (1 = *Not at all effective*, 4 = *Moderately effective,* 7 = *Extremely effective*) as the measure of efficacy beliefs. Participants' moral conviction about their capital punishment or trigger warnings attitudes was measured with the following two items, which were aggregated into a single measure: "To what extent is your position on [capital punishment/trigger warnings] a reflection of your core moral beliefs and convictions?" (1 = *Not at all reflective*, 4 = *Moderately reflective,* 7 = *Extremely reflective*), and "To what extent is your position on [capital punishment/trigger warnings] deeply connected to beliefs about fundamental questions of 'right' and 'wrong'?" (1 = *Not at all connected*, 4 = *Moderately connected,* 7 = *Extremely connected*).

Participants in each study were then randomly assigned to one of two experimental conditions (Condition: blinded or unblinded) in a two-cell between-subjects design. All participants read a brief introduction and methods description of the focal study. For Study 1a, the stimuli read as follows:

> One of the most controversial public issues in recent years has been the effectiveness of capital punishment (the death penalty) in preventing violent crime. Proponents of capital punishment have argued that the possibility of execution deters people who might otherwise commit violent crime, whereas opponents of capital punishment deny this and maintain that the death penalty may even produce violent crime by setting a violent model of behavior. A recent research effort attempted to shed light on this controversy.
>
> Researchers from University of Kansas (Palmer & Crandall, 2007) published a study in

> the *Journal of Social Issues* that looked at the difference in murder rates in states that share a common border but differ in whether their laws permit capital punishment or not. They compiled a list of all possible pairs and then selected 10 pairs of neighboring states that were alike in degree of urbanization (percentage of the population living in metropolitan areas). Using the murder rate (number of willful homicides per 100,000 population) in 2006 as their index, they hypothesized that if capital punishment is effective, the murder rates should be lower in the state with capital punishment laws.
>
> For Study 1b, the stimuli read as follows:
>
> A controversial public question in recent years has been whether trigger warnings reduce the anxiety experienced by people engaging with emotionally arousing content. Proponents of trigger warnings have argued that including these warnings reduces the levels of anxiety felt by individuals presented with difficult or disturbing material, whereas opponents of trigger warnings deny this and believe that trigger warnings do not reduce anxiety and may even *increase* anxiety. A recent research effort attempted to shed light on this controversy.
>
> Researchers from Harvard University (Bellet et al., 2020) published a study in the *Journal of Experimental Psychology: Applied* that looked at the effects of trigger warnings in 462 college students. Participants were randomly assigned to either a control condition or an experimental condition before reading 10 short passages in random order. Five passages contained neutral content, and the other five contained distressing content. In the experimental condition, participants were presented with a trigger warning before reading the distressing passages, while participants in the control condition did not receive any warnings before reading the distressing passages. All participants rated how anxious they felt after reading each passage using slider bars ranging from 0 (not at all) to 100 (very much). The researchers hypothesized that if trigger warnings are effective, then students' average anxiety response to the distressing passages should be lower in the trigger warning condition than in the control condition.

Participants in the unblinded conditions were only required to stay on the page containing this description for 30 seconds before they could proceed to the next page. Alternatively, those in the blinded conditions were required to stay on the page for 30 seconds but were also presented the study quality items immediately after the methods description. The wording of the study quality items for Study 1a are shown in Table 1.1, and the items for Study 2b are shown in Table 1.2. Thus, participants in the blinded conditions evaluated the methodological quality of the studies after reading about the research design but before they knew the results. Importantly,

using more specific measures of study quality limited the plausibility of normative, accuracy-motivated counterexplanations for any observed condition differences these evaluations.

Table 1.1. *Study Quality Items for Study 1a.*

1. How well do you think the researchers measured each states' levels of violent crime? (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)
2. How well does a state's murder rate measure the overall level of violent crime in that state? (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)
3. How appropriate was the sample size (10 pairs of neighboring states) for the study the researchers conducted? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
4. How appropriate was it to analyze these data (neighboring states with different capital punishment laws) to address the research question? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
5. How appropriate was the correlational approach used by the researchers in this study for answering the research question? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
6. How appropriate was it to control for differences in states' degree of urbanization in this study? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
7. Are the data from this study helpful for answering the research question? (1 = *Definitely unhelpful,* 4 = *Neither helpful nor unhelpful,* 7 = *Definitely helpful*)
8. On the scale below, please indicate how valid you would find the results of the above study. (1 = *Very invalid,* 4 = *Neither valid nor invalid,* 7 = *Very valid*)

Table 1.2. *Study Quality Items for Study 1b.*

1. How well do you think the measured participants' reported anxiety?
   (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)

2. How well does participants' reported anxiety measure their actual levels of experienced anxiety?
   (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)

3. How appropriate was the sample size (462 people) for the study the researchers conducted?
   (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)

4. How appropriate was it to analyze these data (college students' responses to written passages) to address the research question?
   (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)

5. How appropriate was the experimental approach used by the researchers in this study for answering the research question?
   (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)

6. How appropriate of a control group did the researchers use in this study?
   (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)

7. Are the data from this study helpful for answering the research question?
   (1 = *Definitely unhelpful,* 4 = *Neither helpful nor unhelpful,* 7 = *Definitely helpful*)

8. On the scale below, please indicate how valid you would find the results of the above study.
   (1 = *Very invalid,* 4 = *Neither valid nor invalid,* 7 = *Very valid*)

On the next page, participants were presented with the full write-up of the appropriate study, which included the introduction and methods descriptions and a brief description of the results. All participants were required to stay on this page for at least 30 seconds before proceeding. In Study 1a, participants read, "Their results, as shown in the table and graph below, were that the murder rates were lower in the state with capital punishment laws than in the state without capital punishment laws for eight of the 10 pairs of states selected for their study. The researchers concluded that the existence of the death penalty does work to deter violent crime." These sentences were accompanied by a table and figure illustrating the results, which are presented in Appendix A. In Study 1b, participants read, "Their results, as shown in the graph below, were that students in the trigger warning condition reported *more* anxiety in response to reading the distressing passages than individuals in the control condition. The researchers concluded that trigger warnings do not work to reduce anxiety and that trigger warnings may actually increase anxiety." This description was accompanied by a corresponding figure, also presented in Appendix A. Participants in the unblinded conditions read the results descriptions and then completed the study quality measures, while participants in the blinded conditions were not presented any additional items before proceeding to the next page.

Then, participants in both conditions were presented with a manipulation check and 10 items assessing their credibility impressions of the study they read about. For the manipulation check, participants responded to the question "What did the results of the study you read about show?" on a 7-pt. scale. The response options for Study 1a (1 = *Capital punishment is not effective at all*, 4 = *Capital punishment is moderately effective*, 7 = *Capital punishment is extremely effective*) and Study 1b (1 = *Trigger warnings dramatically decrease anxiety*, 4 = *Trigger warnings do not increase or decrease anxiety*, 7 = *Trigger warnings dramatically*

increase anxiety) differed to match the corresponding materials. The credibility impression items, presented in Table 1.3, were nearly identical across studies, only differing in references to specific information about the relevant study (e.g., the university where the research was conducted). These items captured various perceptions of the study that I anticipated may be influenced by both participants' prior beliefs and their methodological quality evaluations.

Table 1.3. *Credibility impression Items for Studies 1a and 1b.*

1. How well done (or poorly done) was this study?
   (1 = *Very poorly done*, 4 = *Neither well nor poorly done,* 7 = *Very well done*)

2. How high-quality (or low-quality) was this study?
   (1 = *Extremely low-quality study,* 4 = *Neither high-quality nor low-quality,* 7 = *Extremely high-quality study*)

3. How trustworthy (or untrustworthy) are the researchers who conducted this study?
   (1 = *Extremely untrustworthy,* 4 = *Neither trustworthy nor untrustworthy,* 7 = *Extremely trustworthy*)

4. How objective (or biased) do you think the researchers were when conducting their study?
   (1 = *Researchers were very biased*, 4 = *Researchers were neither objective nor biased,* 7 = *Researchers were very objective)*

5. How believable (or not believable) are the results of this study?
   (1 = *Not believable at all,* 4 = *Moderately believable,* 7 = *Extremely believable*)

6. How credible (or not credible) are the results of this study?
   (1 = *Not credible at all,* 4 = *Moderately credible,* 7 = *Extremely credible*)

7. How convincing (or unconvincing) is this study as evidence of the **effectiveness of capital punishment/ ineffectiveness of trigger warnings**?
   (1 = *Completely unconvincing,* 4 = *Neither convincing nor unconvincing,* 7 = *Completely convincing*)

8. How reputable (or disreputable) is the **University of Kansas/Harvard University**, the place where this research was conducted?
   (1 = *Very disreputable*, 4 = *Neither reputable nor disreputable,* 7 = *Very reputable*)

9. How reputable (or disreputable) is the **Journal of Social Issues/Journal of Experimental Psychology: Applied**, the journal where this research was published?
   (1 = *Very disreputable*, 4 = *Neither reputable nor disreputable,* 7 = *Very reputable*)

10. How strong (or weak) of evidence is this study for the **effectiveness of capital punishment as a deterrent of violent crime/ineffectiveness of trigger warnings in reducing anxiety**?
   (1 = *Extremely weak evidence,* 4 = *Neither strong nor weak evidence,* 7 = *Extremely strong evidence*)

Note: Bolded phrases differed between Studies 1a and 1b.

On the following page, participants were asked to complete policy support and efficacy items that were nearly identical to those presented to them at the beginning of the respective studies. For these measures, participants were asked, "After having read about this study, to what extent do you *now* support or oppose capital punishment in the United States/the use of trigger warnings]?" (1 = *Strongly oppose*, 4 = *Neither support nor oppose*, 7 = *Strongly support*), and "After having read about this study, to what extent do you now believe that [capital punishment is an effective deterrent of violent crime/trigger warnings are effective at reducing anxiety for people exposed to distressing content]?" (1 = *Not at all effective*, 4 = *Moderately effective*, 7 = *Extremely effective*). These served as the second time point for the belief measures, from which measurements of actual belief change (i.e., actual attitude change, Miller et al., 1993) were ultimately constructed (Time 2 beliefs – Time 1 beliefs). Additionally, participants were asked to complete items of perceived attitude change for their policy support and efficacy beliefs. These items read, "Compared to before you read about this study, how has your support or opposition to capital [punishment/trigger warnings] changed?" (1 = *Much more opposed to* [*capital punishment/trigger warnings]*, 4 = *Not changed*, 7 = *Much more opposed to* [*capital punishment/trigger warnings*]), and "Compared to before you read about this study, how have your views on the effectiveness of [capital punishment/trigger warnings] changed?" (1 = *Much less effective than I thought before*, 4 = *Not changed*, 7 = *Much more effective than I thought before*).

Before concluding the study, participants completed demographic information about their age, sex, ethnicity, income, education, political orientation (social and economic, separately), and political party affiliation (these items are presented in Appendix A). Finally, participants were

given the opportunity to share any thoughts or feelings about the study in an open-response question and presented with a debriefing before concluding their participation.

Results

Following my preregistration, I excluded participants who did not pass our English comprehension checks, failed our manipulation check (responding 1 = "*Capital punishment is not effective at all*" in Study 1a or below 4 = "*Trigger warnings do not increase or decrease anxiety"* in Study 1b), or finished the survey in less than three minutes. This resulted in a sample of 466 participants for the confirmatory analyses in Study 1a and 463 in Study 1b (including the full sample did not substantively alter the results). The distribution of participants across the two experimental conditions was roughly equivalent in both Study 1a ($n_{blinded}$ = 234, $n_{unblinded}$ = 232) and 1b ($n_{blinded}$ = 235, $n_{unblinded}$ = 228).

In Study 1a, participants overall were slightly opposed to capital punishment (*M* = 3.43, *SD* = 1.95) and thought that capital punishment is somewhat effective at deterring violent crime (*M* = 3.19, *SD* = 1.86). These measures were strongly correlated (*r* = 0.72). In Study 1b, participants were supportive of trigger warnings (*M* = 5.30, *SD* = 1.54) and thought that trigger warnings are moderately effective at reducing anxiety (*M* = 4.46, *SD* = 1.60). The two prior belief measures were strongly correlated in this study as well (*r* = 0.62). In both studies, participants were moderately morally convicted[1] about the topic, were slightly liberal in their social and economic political orientation on average, and leaned Democratic (details are presented in Appendix A). The z-score indices of skewness and kurtosis were less than ±2 for the

[1] While I preregistered analyses using moral conviction as part of three-way interactions, these studies were underpowered to reliably test those analyses. Thus, I do not discuss the moral convictions items further in Studies 1a and 1b, but I present more highly-powered analyses that include moral conviction in Study 2.

key variables in each study, indicating that the distributions of responses were sufficiently normal for parametric analyses (Field et al., 2012).

Study quality evaluations

The 10 study quality items had high internal reliability in both Study 1a (Cronbach's α = 0.89) and 1b (Cronbach's α = 0.89) and thus were averaged into composite measures in each study[2]. Despite being conceptually independent, participants' evaluations of one aspect of the presented study (e.g., sample size) tracked with their evaluations of other aspects of the study (e.g., use of appropriate controls). On average, participants rated the presented study as being of decent quality in both Study 1a (M = 4.81, SD = 1.08) and 1b (M = 5.05, SD = 1.04).

To test my hypotheses regarding to the existence of partisan bias (Research Question 1) and the types of prior beliefs that bias partisans' evaluations (Research Question 2), I conducted a series of linear regression analyses using the GAMLj package in jamovi.

Influence of prior support beliefs

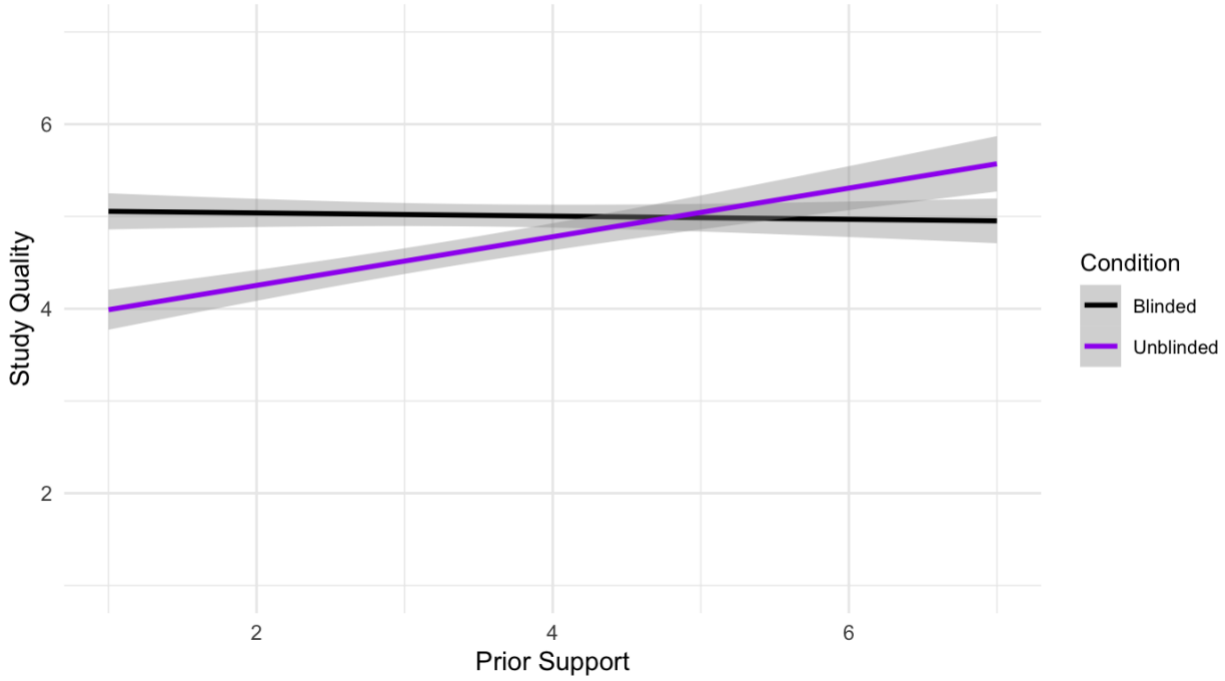
To start, I constructed models predicting study quality evaluations from condition (dummy-coded, 0 = blinded), prior support (mean-centered), and their interaction.

Study 1a. Omnibus tests showed a significant main effect of condition $F(1, 462) = 17.17$, $p < .001$, $\eta_p^2 = .04$, such that participants in the blinded condition provided more positive quality evaluations than did participants in the unblinded condition, but there was no main effect of prior support for capital punishment, $F(1, 462) = 0.27$, $p = .60$, $\eta_p^2 = .00$. This was qualified by a significant interaction between condition and prior support, $F(1, 462) = 34.54$, $p < .001$, $\eta_p^2 = .07$. Simple effects analyses showed that, as hypothesized, prior support was associated with

[2] The internal reliability statistics for the study quality items in the blinded conditions of Study 1a (Cronbach's α = 0.89) and Study 1b (Cronbach's α = 0.87) were essentially equivalent to those in the unblinded conditions of Study 1a (Cronbach's α = 0.90) and Study 1b (Cronbach's α = 0.90).

study quality evaluations in the unblinded condition, b = 0.26, SE = 0.03, p < .001, 95% CI [0.20, 0.33], β = 0.48, but not in the blinded condition, b = -0.02, SE = 0.03, p = .60, 95% CI [-0.08, 0.05], β = -0.03. As illustrated in Figure 1.1, prior support was unrelated to blinded participants' quality evaluations, but prior support was strongly predictive of quality evaluations for unblinded participants, such that having stronger prior support was associated with providing more positive evaluations.

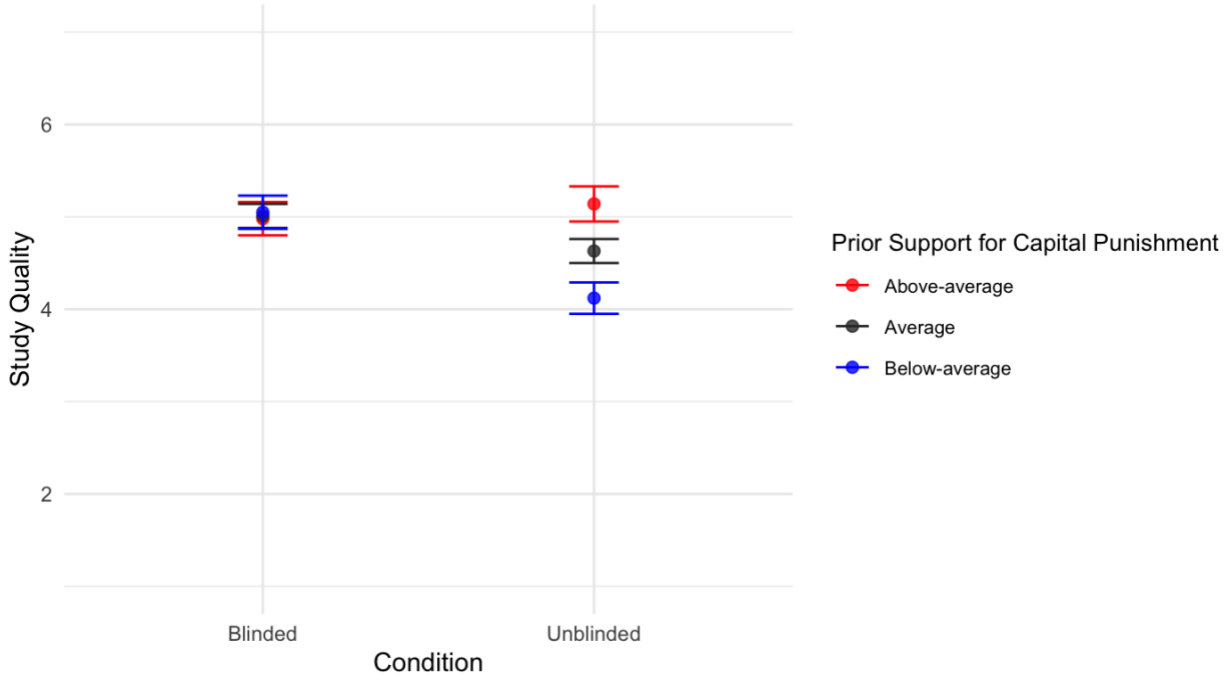
Figure 1.1. Average Study Quality Evaluations by Prior Support and Condition in Study 1a
Error bands represent standard errors



Additionally, an analysis of the estimated marginal means, illustrated in Figure 1.2, provided further support for the existence of partisan bias. There were no significant differences in quality evaluations across levels of prior support for participants in the blinded condition, yet there were significant differences (as indicated by nonoverlapping 95% CIs) across levels of prior belief in the unblinded condition. Moreover, there were significant differences across conditions for those with either average or below-average (M – 1SD) prior support for capital

punishment—for whom the presented results were counter-attitudinal—but not for those with above-average (*M* + 1*SD*) prior support. Unblinded participants who came into the study more opposed to capital punishment had significantly lower quality evaluations in than their blinded counterparts, yet this bias in quality evaluations did not emerge in participants for whom the results aligned with their prior support.

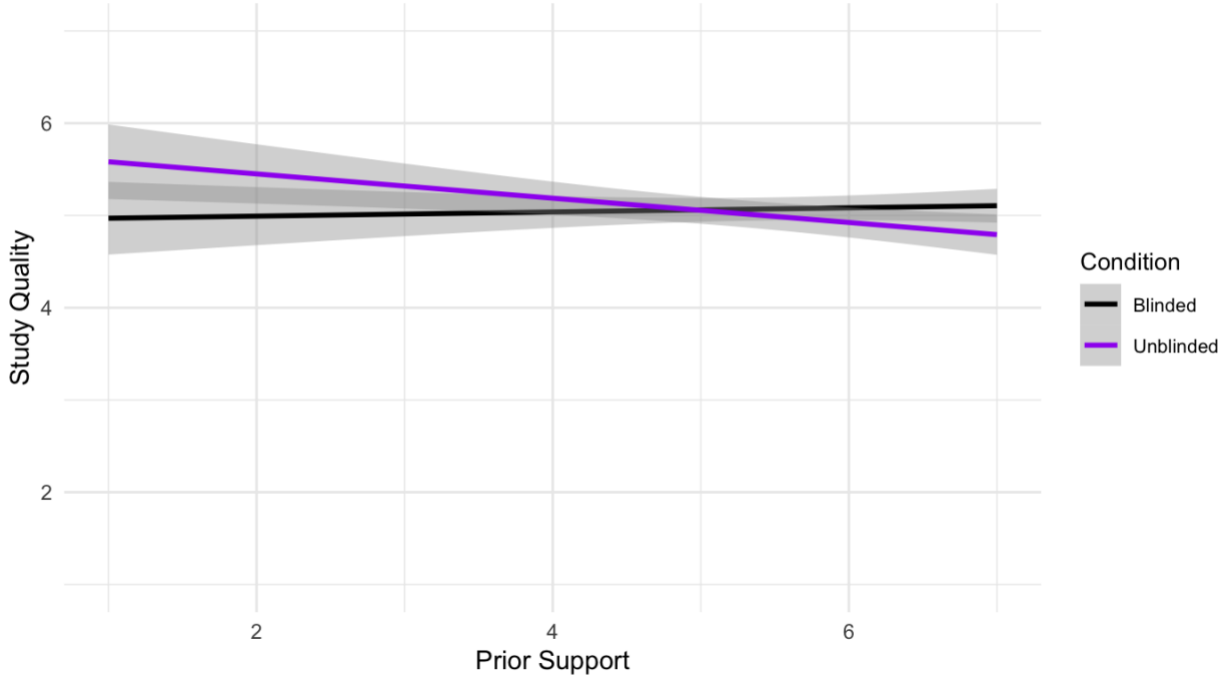
Figure 1.2. Average Study Quality Evaluations by Condition and Prior Support in Study 1a
Error bars represent 95% CIs



Study 1b. In Study 1b, there was neither a significant main effect of condition $F(1, 459) = 0.28$, $p = .60$, $\eta_p^2 = .00$, nor prior support for trigger warnings, $F(1, 459) = 0.24$, $p = .63$, $\eta_p^2 = .00$. Yet as in Study 1a, there was a significant interaction between condition and prior support, $F(1, 459) = 5.98$, $p = .015$, $\eta_p^2 = .01$. Once again, simple effects analyses showed that, consistent with my hypotheses, prior support was associated with study quality evaluations in the unblinded condition, $b = -0.13$, $SE = 0.04$, $p = .002$, 95% CI [-0.22, -0.05], $\beta = -0.19$, but not in the blinded condition, $b = 0.02$, $SE = 0.05$, $p = .63$, 95% CI [-0.07, 0.11], $\beta = 0.03$. Blinded participants'

prior support was inhibited from influencing their quality evaluations; in contrast, the more that unblinded participants supported trigger warnings before reading the presented materials, the more negatively they evaluated a study that found trigger warnings to be ineffective. This interaction, illustrated in Figure 1.3, was smaller than the one observed in Study 1a.

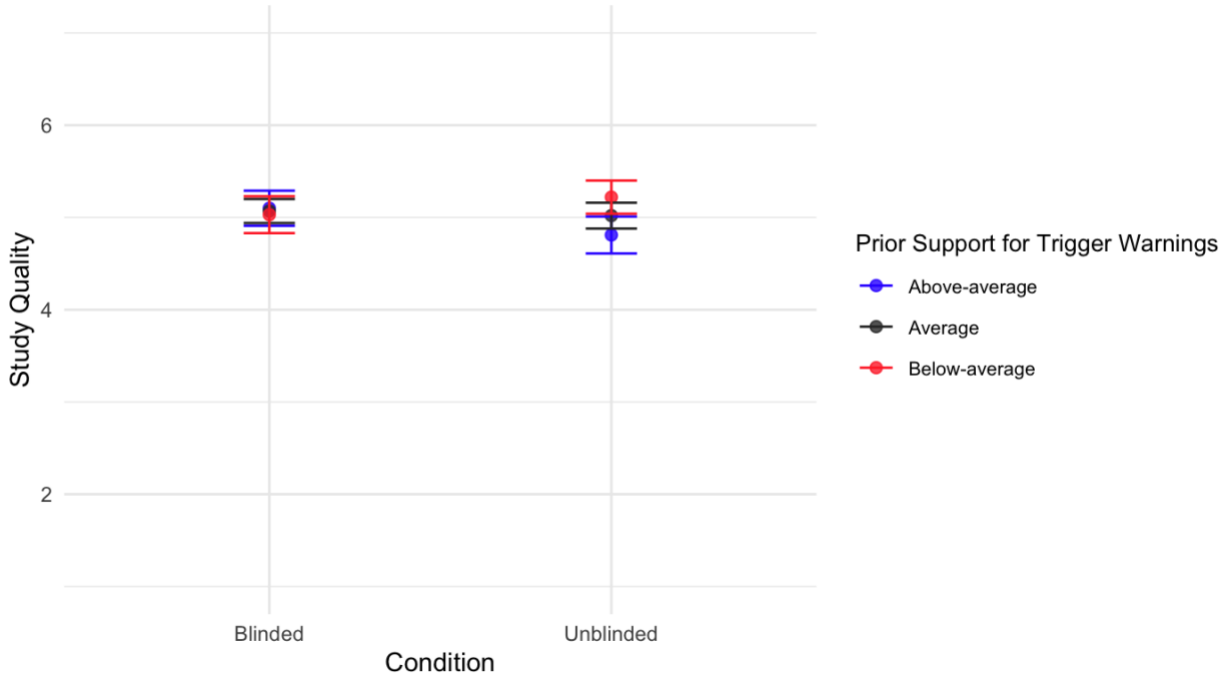
Figure 1.3. Average Study Quality Evaluations by Prior Support and Condition in Study 1b
Error bands represent standard errors



An analysis of the estimated marginal mans, illustrated in Figure 1.4, yielded a similar pattern of results as Study 1a. Participants for whom the presented results were more politically-unfriendly, those with above-average prior support, evaluated the study slightly more negatively in the unblinded condition. Additionally, participants for whom the presented results were more politically-friendly, those with below-average prior support, evaluated the study slightly more positively in the unblinded condition. However, while these results were like those of Study 1a, the differences between conditions were not statistically significant in Study 1b.

Figure 1.4. Average Study Quality Evaluations by Condition and Prior Support in Study 1b

Error bars represent 95% CIs



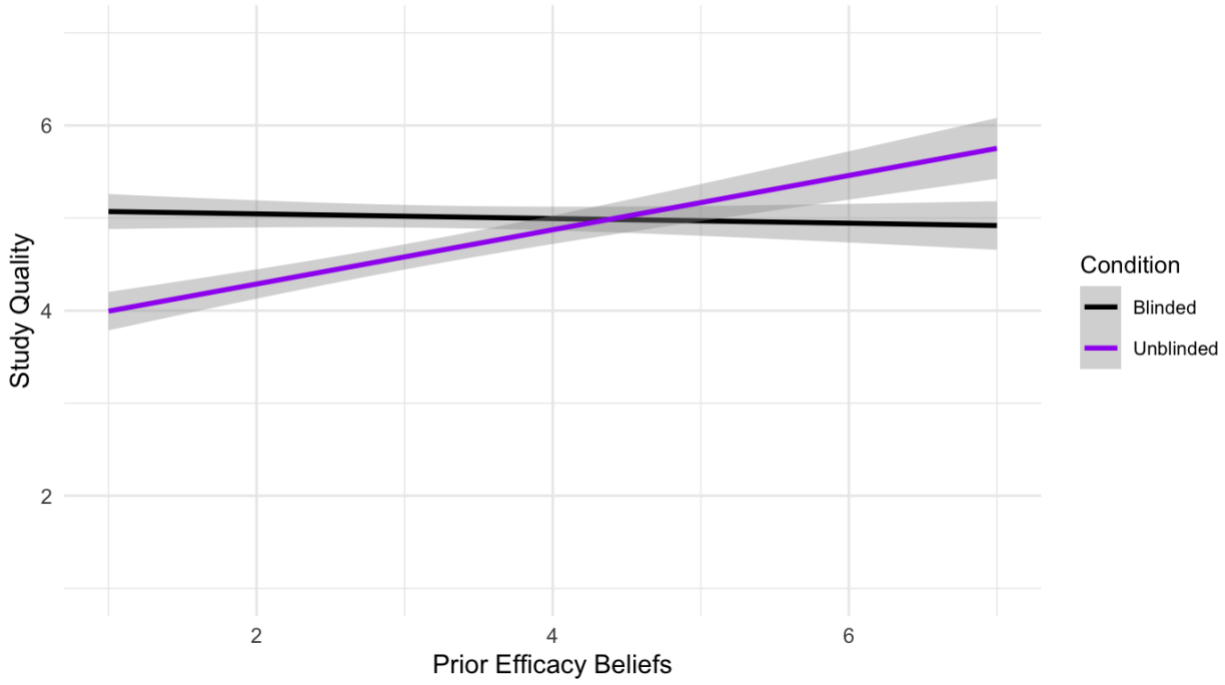
To summarize, study quality evaluations were significantly influenced by participants' prior support in the unblinded conditions, but not in the blinded conditions. If participants were making purely accuracy-motivated judgments in the unblinded conditions, then participants' prior support should have exerted the same influence on their quality evaluations across conditions. Yet I observed that these beliefs had a significantly stronger effect on the quality evaluations of participants in the unblinded condition, reflecting directionally biased evaluations compared to the blinded comparisons. Moreover, participants for whom the presented results were politically-unfriendly (below-average prior support in Study 1a and above-average prior support in Study 1b) had greater differences between their blinded and unblinded evaluations, although the differences between conditions for these participants were only significant in Study 1a.

Influence of prior efficacy beliefs

Next, I constructed similar models predicting study quality evaluations from condition (dummy-coded, 0 = blinded), prior efficacy beliefs (mean-centered), and their interaction.

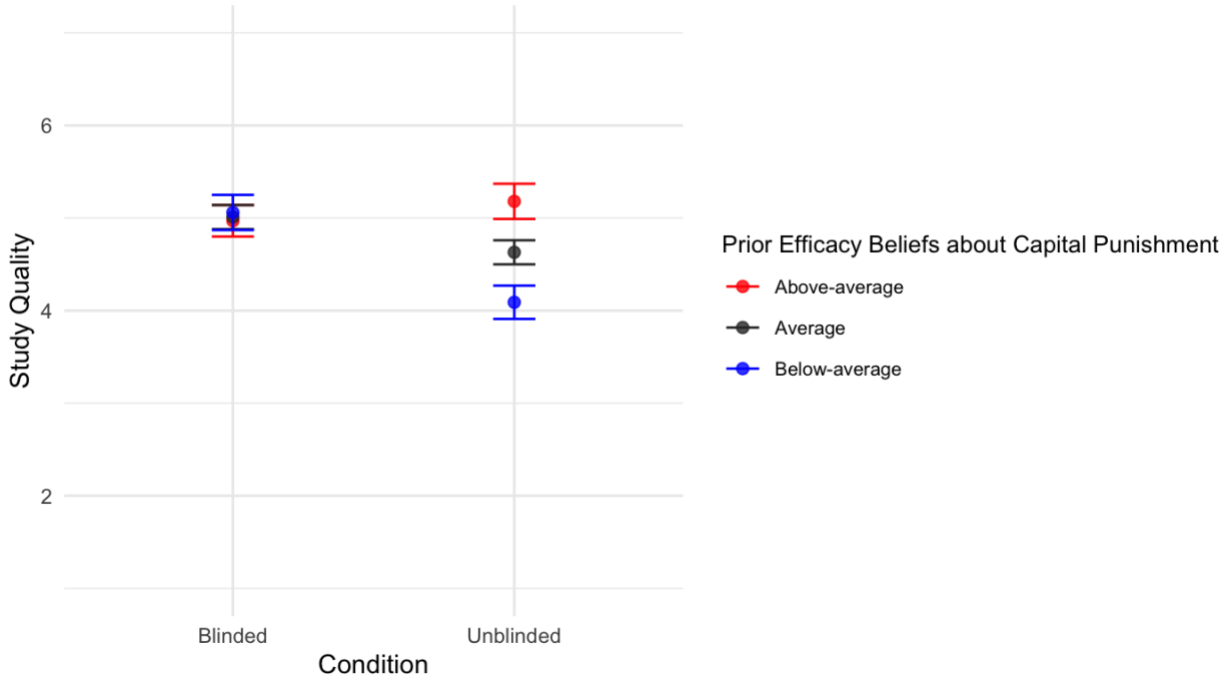
Study 1a. Omnibus tests for Study 1a showed a significant main effect of condition $F(1, 462) = 17.17$, $p < .001$, $\eta_p^2 = .04$, but not of prior efficacy beliefs about capital punishment, $F(1, 462) = 0.55$, $p = .46$, $\eta_p^2 = .00$. As hypothesized, the condition main effect was qualified by a significant interaction, $F(1, 462) = 40.68$, $p < .001$, $\eta_p^2 = .08$. Simple effects analyses showed that believing capital punishment is effective was associated with positive study quality evaluations in the unblinded condition, $b = 0.29$, $SE = 0.04$, $p < .001$, 95% CI [0.22, 0.36], $\beta = 0.50$, but not in the blinded condition, $b = -0.03$, $SE = 0.03$, $p = .46$, 95% CI [-0.09, 0.04], $\beta = -0.04$. As illustrated in Figure 1.5, unblinded participants' prior efficacy beliefs directionally biased their quality evaluations relative to the blinded evaluations of their partisan counterparts.

Figure 1.5 Average Study Quality Evaluations by Prior Efficacy Beliefs and Condition in Study 1a
Error bands represent standard errors



Moreover, an analysis of the estimated marginal means showed that, as was the case in the analysis of prior support in Study 1a, this bias was driven by participants for whom the presented results were politically-unfriendly. That is, there were significant differences (as indicated by nonoverlapping 95% CIs) across conditions for those with either average or below-average prior efficacy beliefs about capital punishment, but this was not the case for those with above-average prior efficacy beliefs. These estimated marginal means are shown in Figure 1.6.

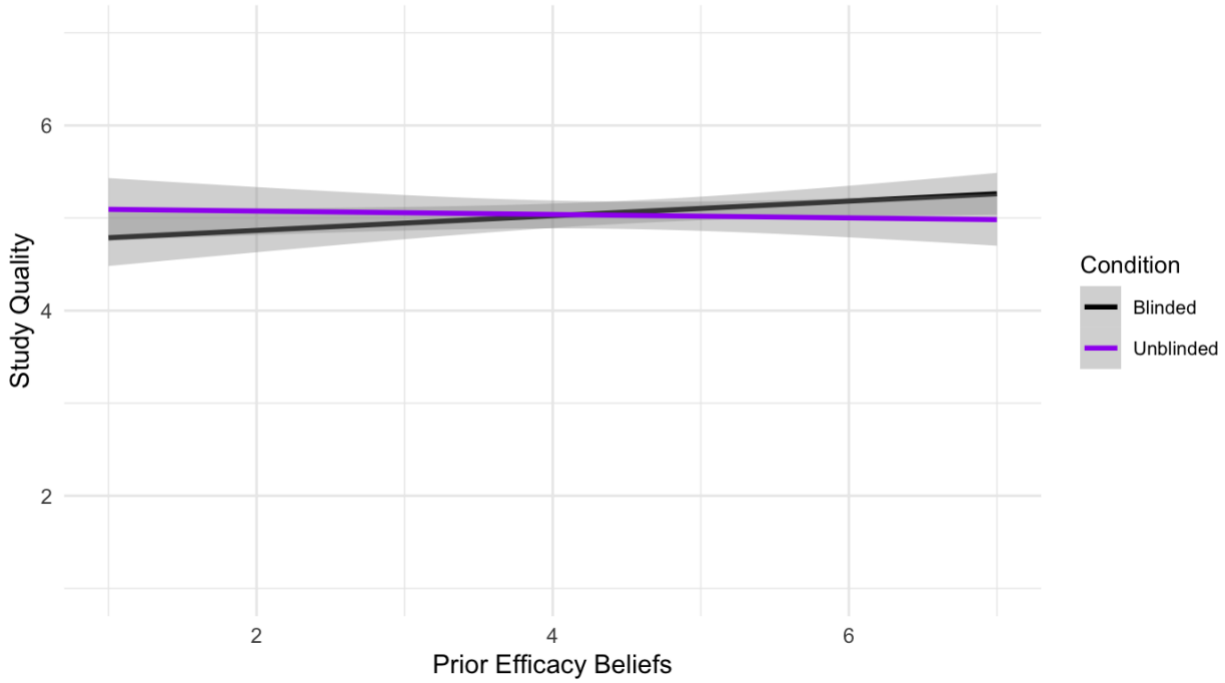
Figure 1.6. Average Study Quality Evaluations by Condition and Prior Efficacy Beliefs in Study 1a
Error bars represent 95% CIs



Study 1b. In contrast to the results of Study 1a, in Study 1b, there was not a significant main effect of condition $F(1, 462) = 0.10$, $p = .75$, $\eta_p^2 = .00$, prior efficacy beliefs about trigger warnings, $F(1, 462) = 3.34$, $p = .07$, $\eta_p^2 = .01$, nor a significant interaction, $F(1, 462) = 2.58$, $p = .11$, $\eta_p^2 = .01$. These results, illustrated in Figure X, indicated that prior efficacy beliefs did not influence participants' quality evaluations in either condition and, thus, did not bias participants unblinded quality evaluations.

Figure 1.7. Average Study Quality Evaluations by Prior Efficacy Beliefs and Condition in Study 1b

Error bands represent standard errors



Thus, the predicted interaction between prior efficacy beliefs and condition emerged in Study 1a but not in Study 1b. Participants' prior beliefs about the efficacy of capital punishment biased their unblinded quality evaluations in Study 1a, but participants' prior beliefs about the efficacy of trigger warnings did not significantly bias their quality evaluations in Study 1b.

Relative influence of prior support and prior efficacy beliefs.

To further address Research Question 2, I explored the relative influence of prior support and prior efficacy beliefs on quality evaluations. I constructed regression models using the condition variable, both prior belief measures, and the two-way interactions between condition and each prior belief variable. The key terms in these models are the interaction terms, which indicate the influence that each prior belief measure had on quality evaluation when accounting for the shared variance between the two measures.

In Study 1a, the main effect of condition was significant, $F(1, 460) = 16.90$, $p < .001$, $\eta_p^2 = .04$, but neither the main effect of prior support, $F(1, 460) = 0.00$, $p = .99$, $\eta_p^2 = .00$, nor the

main effect of prior efficacy beliefs, $F(1, 460) = 0.27$, $p = .60$, $\eta_p^2 = .00$, were significant. Nevertheless, the interaction between prior efficacy beliefs and condition was significant and in the same direction as the previous analyses, $F(1, 460) = 8.72$, $p = .003$, $\eta_p^2 = .02$. The interaction between prior support and condition was not significant, $F(1, 460) = 3.47$, $p = .063$, $\eta_p^2 = .01$, though it was also trending in the direction of the previous analyses. Prior efficacy beliefs thus significantly biased participants' quality evaluations above and beyond the shared influence that prior support and prior efficacy beliefs had on participants' judgments in the unblinded condition.

A different pattern of effects emerged in Study 1b. The main effect of condition was not significant, $F(1, 457) = 0.20$, $p = .66$, $\eta_p^2 = .00$, and neither was the main effect of prior support, $F(1, 457) = 0.37$, $p = .55$, $\eta_p^2 = .00$, nor the main effect of prior efficacy beliefs, $F(1, 457) = 3.56$, $p = .06$, $\eta_p^2 = .01$. The interaction between prior efficacy beliefs and condition was not significant either, $F(1, 457) = 0.21$, $p = .65$, $\eta_p^2 = .00$. However, the interaction between prior support and condition was significant in this analysis, $F(1, 457) = 5.62$, $p = .018$, $\eta_p^2 = .01$, indicating that supporting trigger warnings was associated with mor negative study quality evaluations in the unblinded condition, $b = -0.22$, $SE = 0.06$, $p < .001$, 95% CI [-0.34, -0.11], $\beta = -0.33$, but not in the blinded condition, $b = -0.03$, $SE = 0.05$, $p = .55$, 95% CI [-0.14, 0.07], $\beta = -0.05$. This result indicated that only participants' prior support biased their quality evaluations in Study 1b.

Exploratory analysis of political orientation and party affiliation.

Appendix A presents exploratory analyses using political orientation and party affiliation, respectively, as predictors of study quality evaluations. In both studies, more-liberal-leaning participants had lower quality evaluations in the unblinded conditions, but participants' political

orientation and party affiliation were not predictive of quality evaluations in the blinded conditions. However, the differences between the blinded and unblinded estimates were only significant in Study 1a. These results mirrored those from the models including prior support to predict quality evaluations, further suggesting that unblinded participants' political motivations exerted a biasing influence on their evaluations.

Credibility impressions

The credibility impressions measures had high internal reliability in both Study 1a (Cronbach's $\alpha = 0.92$) and 1b (Cronbach's $\alpha = 0.91$). As with the study quality items, this indicated that participants' impressions of one aspect of the study (e.g., how believable the results were) were strongly correlated with their perceptions of other, conceptually unrelated aspects of the study (e.g., how reputable the university was at which the work was conducted). Given this high internal reliability, I made composite credibility impressions items for both Study 1a and 1b.

Participants had modestly positive credibility impressions on average in Study 1a (M = 4.58, SD = 1.03) and 1b (M = 5.02, SD = 1.00). In Study 1a, participants in the blinded condition (M = 4.67, SD = 0.95) did not have significantly more positive credibility impressions than those in the unblinded condition (M = 4.50, SD = 1.11), Welch's $t(451.65) = 1.72$, $p = .086$, Cohen's d = 0.16. This was also the case in Study 1b, where participants in the blinded condition (M = 5.06, SD = 0.95) had roughly equivalent credibility impressions as participants in the unblinded condition (M = 4.99, SD = 1.04), Welch's $t(453.64) = 0.77$, $p = .44$, Cohen's d = 0.07. Without accounting for participants' prior beliefs, blinding participants' quality evaluations did not appear to meaningfully influence their credibility impressions of the study.

However, to address Research Question 3 more rigorously, I conducted a series of analyses to assess whether the effect of blinding on credibility impressions may have varied for participants with different prior beliefs. I also explored whether condition differences in study quality evaluations could help explain any such effects.

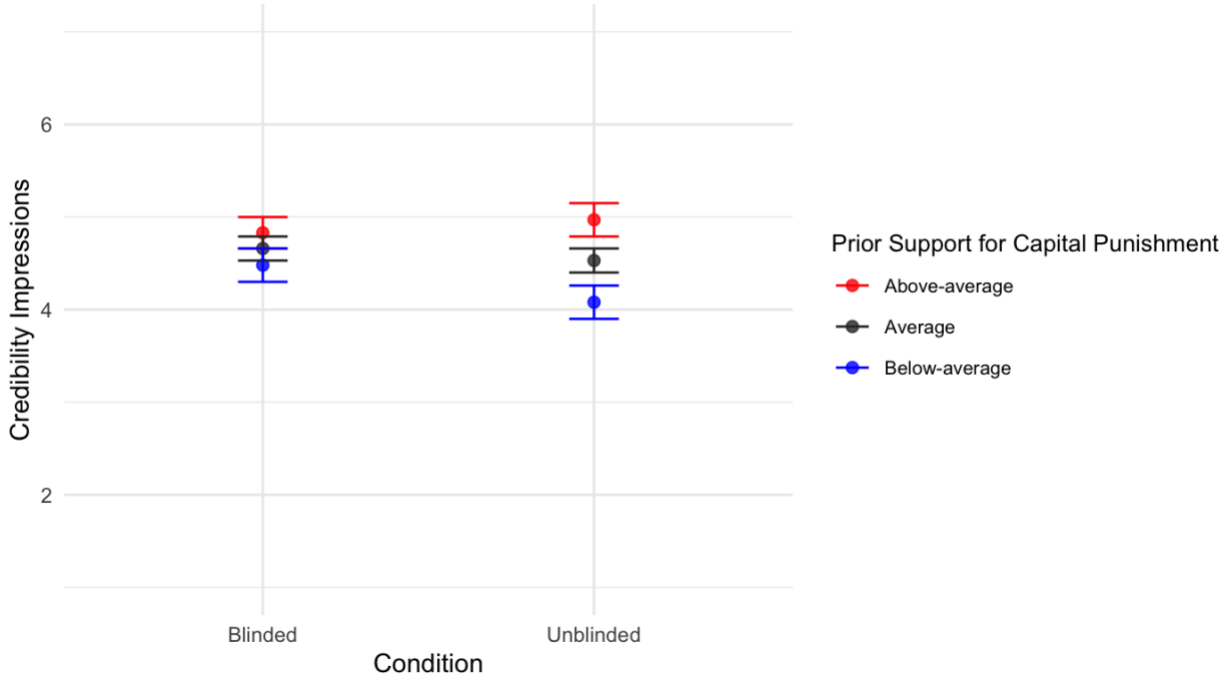
Influence of prior support.

Study 1a. To start, I constructed linear regression models predicting credibility impressions from condition (dummy-coded, 0 = blinded), prior support (mean-centered), and their interaction. In Study 1a, there was not a significant main effect of condition, $F(1, 462) = 2.01$, $p = .16$, $\eta_p^2 = .00$, but there was a main effect of prior support $F(1, 462) = 46.61$, $p < .001$, $\eta_p^2 = .09$, qualified by a significant interaction, $F(1, 462) = 9.05$, $p = .003$, $\eta_p^2 = .02$. While greater support for capital punishment was predictive of more positive credibility impressions across conditions, this association was significantly stronger in the unblinded condition, $b = 0.22$, $SE = 0.03$, $p < .001$, 95% CI [0.16, 0.30], $\beta = 0.43$, than in the blinded condition, $b = 0.09$, $SE = 0.03$, $p = .003$, 95% CI [0.03, 0.15], $\beta = 0.17$.

Moreover, an analysis of the estimated marginal means, illustrated in Figure 1.8, showed that participants who were more opposed to capital punishment (below-average support, $M - 1SD$) had significantly more positive credibility impressions of the study in the blinded condition compared to the unblinded condition. Blinding the quality evaluations of participants for whom the presented results were most politically-unfriendly increased their overall impressions of the credibility of the presented evidence. Those with average and above-average prior support, in contrast, did not significantly differ in their credibility impressions across conditions. Thus, the blinding manipulation increased the credibility impressions of participants for whom the

presented results were politically-unfriendly, but it did not have a significant effect on the credibility impressions of participants with other levels of support for capital punishment.

Figure 1.8. Average Credibility Impressions by Condition and Prior Support in Study 1a
Error bars represent 95% CIs



To explore whether study quality evaluations could help explain the differences in credibility impressions between conditions for participants with below-average support for capital punishment, I constructed moderated mediation models for each study using the jAMM package in jamovi. These models included condition (dummy-coded, 0 = blinded) as the predictor, study quality as the mediator, prior support (mean-centered) as the moderator, and credibility impressions as the outcome variable. This model, illustrated in Figure 1.9, allowed me to assess the direct effect that condition had on credibility impressions, the indirect effect that condition had on credibility impression through evaluations of study quality, and whether these direct and indirect effects differed by participants' prior support for capital punishment. Confidence intervals for these model estimates were calculated using 1000 bootstrap replications.

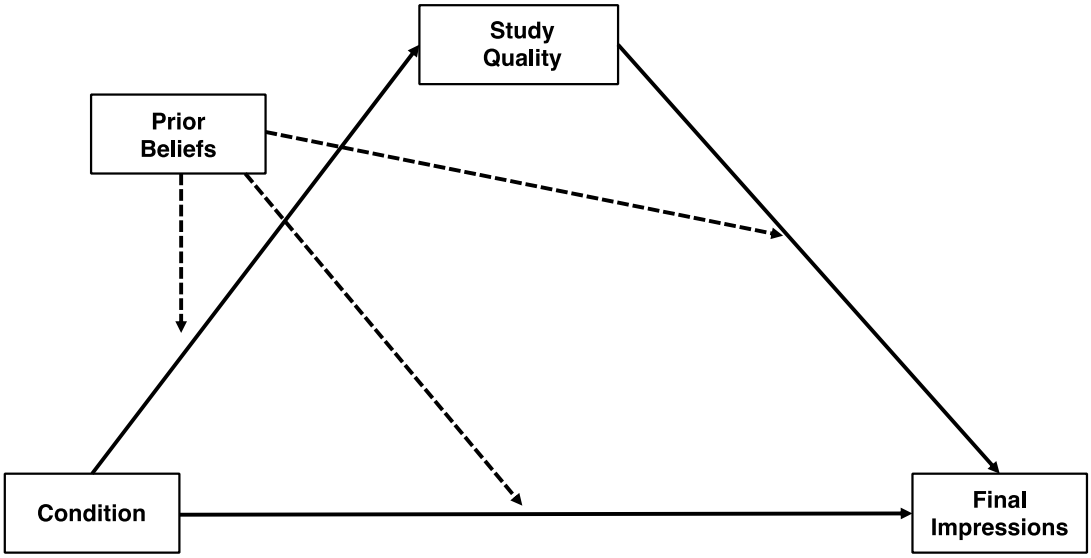


Figure 1.9. Illustration of the moderated mediation model predicting credibility impressions.

The full set of model estimates are presented in Table 1.4. Starting with participants with stronger support for capital punishment (above-average support, $M + 1SD$), there was not a significant indirect effect through evaluations of study quality, a direct effect of condition of credibility impressions, nor a total effect of condition. This indicated that, across conditions, participants who read politically-friendly results were not significantly influenced by their prior support when evaluating study quality or forming credibility impressions in Study 1a. This was not the case for participants with average prior support or who were more strongly opposed to capital punishment (below-average support, $M - 1SD$). For participants with average and below-average prior support, there were significant indirect effects, such that being in the unblinded condition caused these participants to provide more negative quality evaluations than their attitudinally-equivalent counterparts in the blinded condition, which subsequently led the unblinded participants to form more negative credibility impressions of the study. However, there were countervailing direct effects of condition on credibility impressions for both groups of participants, although this direct effect was only significant for participants with below-average

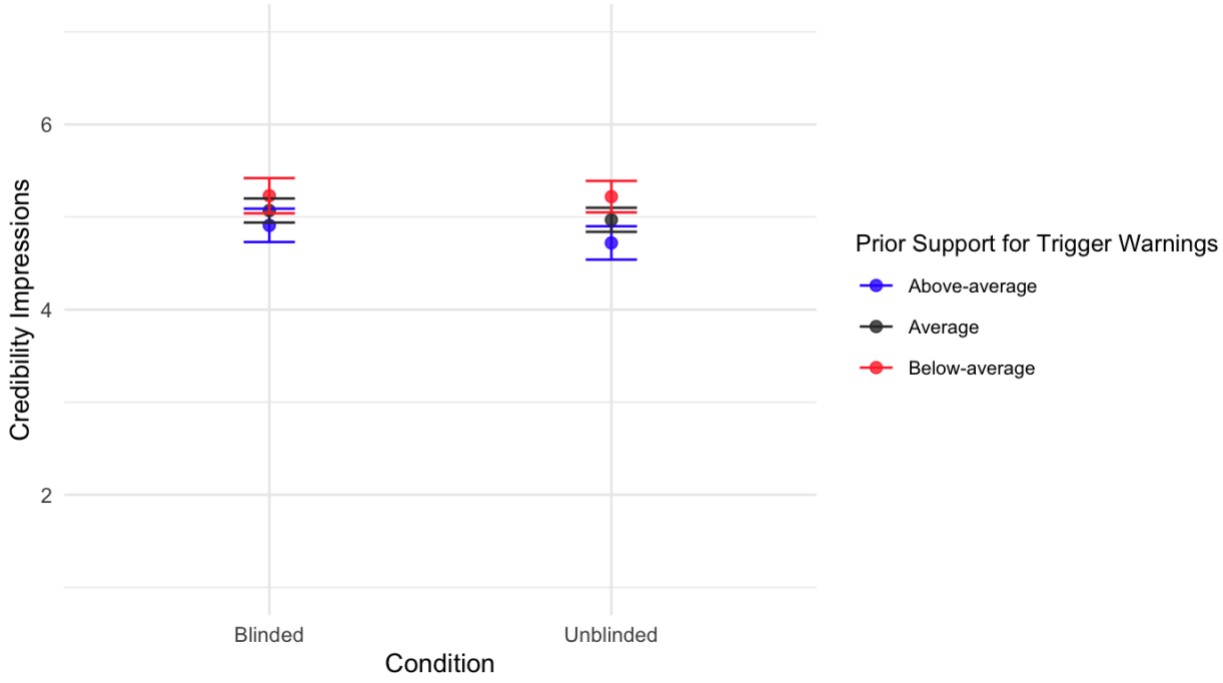
prior support. That is, being in the unblinded condition directly increased credibility impressions for those with average and below-average prior support, even though it also indirectly *decreased* credibility impressions by lowering these participants' evaluations of study quality. Nevertheless, the indirect effects were roughly three times the magnitude of the positive direct effects, resulting in negative total effects for both groups of participants (although the total effect was only significant for participants with below-average prior beliefs).

Ultimately, the impact that the blinding manipulation had on participants' credibility impressions in Study 1a depended on participants' prior support for capital punishment. Participants who were more opposed to capital punishment—and thus read politically-unfriendly results—had more negative credibility impressions in the unblinded condition, when they evaluated the study's quality while knowing its results. This effect on credibility impressions was explained by the influence of the biased quality evaluations that these unblinded participants provided. Put differently, blinding participants quality evaluations caused them to form more positive credibility impressions of politically-unfriendly information by elevating their quality evaluations of that information.

Study 1b. A similar, but not fully consistent, pattern of results emerged in Study 1b. Like in the Study 1a regression model, there was not a significant main effect of condition, $F(1, 459) = 1.27$, $p = .26$, $\eta_p^2 = .00$, but there was a main effect of prior support $F(1, 549) = 5.70$, $p = .017$, $\eta_p^2 = .01$. However, the interaction effect was not significant in Study 1b, $F(1, 459) = 0.95$, $p = .33$, $\eta_p^2 = .00$. As illustrated in Figure 1.9, participants with above-average support for trigger warnings (for whom the presented results were politically-unfriendly) had slightly more negative credibility impressions in the unblinded condition than in the blinded condition, but this difference was not significant. Those with average and below-average prior support did not

significantly differ in their credibility impressions across conditions. The results of the moderated mediation analysis, presented in Table 1.5, also showed a similar pattern to the analysis conducted for Study 1a. Specifically, there was an indirect effect for participants with above-average prior beliefs that was trending toward significance, such that these participants tended to have lower study quality evaluations and, subsequently, more negative credibility impressions in the unblinded condition. However, these estimates were not particularly precise, suggesting a lack of statistical power to detect the indirect effect. Those with average and below-average prior support did not have significant total or indirect effects of condition on credibility impressions.

Figure 1.9. Average Credibility Impressions by Condition and Prior Support in Study 1b
Error bars represent 95% CIs



In sum, Study 1b provided weak evidence of the effects observed more clearly in Study 1a. Participants for whom the presented results about trigger warnings were politically-unfriendly provided slightly lower evaluations of study quality when they made unblinded

quality evaluations, and this caused them to have nonsignificantly more negative credibility impressions than those with similar attitudes who made blinded quality evaluations.

Table 1.4. *Moderated mediation estimates of condition predicting credibility impressions by prior support for Study 1a.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.61	0.11	[-0.82, -0.41]	-0.31	< .001
		Condition ⇒ Study Quality	-0.93	0.14	[-1.20, -0.66]	-0.43	< .001
		Study Quality ⇒ Credibility impressions	0.66	0.04	[0.58, 0.74]	0.72	< .001
	Direct	Condition ⇒ Credibility impressions	0.21	0.09	[0.04, 0.40]	0.11	.02
	Total	Condition ⇒ Credibility impressions	-0.40	0.13	[-0.65, -0.15]	-0.19	.002
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.28	0.07	[-0.43, -0.14]	-0.14	< .001
		Condition ⇒ Study Quality	-0.38	0.10	[-0.58, -0.20]	-0.18	< .001
		Study Quality ⇒ Credibility impressions	0.74	0.03	[0.68, 0.80]	0.76	< .001
	Direct	Condition ⇒ Credibility impressions	0.11	0.06	[-0.01, 0.23]	0.05	.061
	Total	Condition ⇒ Credibility impressions	-0.13	0.09	[-0.31, 0.05]	-0.06	.16
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.13	0.10	[-0.07, 0.34]	0.06	.20
		Condition ⇒ Study Quality	0.16	0.13	[-0.09, 0.41]	0.08	.20
		Study Quality ⇒ Credibility impressions	0.81	0.04	[0.74, 0.89]	0.80	< .001
	Direct	Condition ⇒ Credibility impressions	0.01	0.09	[-0.16, 0.19]	0.01	.87
	Total	Condition ⇒ Credibility impressions	0.15	0.13	[-0.11, 0.40]	0.07	.26

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 1.5. *Moderated mediation estimates of condition predicting credibility impressions by prior support for Study 1b.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.12	0.08	[-0.04, 0.28]	0.06	.15
		Condition ⇒ Study Quality	0.19	0.13	[-0.07, 0.45]	0.09	.15
		Study Quality ⇒ Credibility impressions	0.63	0.05	[0.54, 0.72]	0.71	< .001
	Direct	Condition ⇒ Credibility impressions	-0.11	0.07	[-0.26, 0.03]	-0.06	.13
	Total	Condition ⇒ Credibility impressions	-0.01	0.13	[-0.27, 0.24]	-0.01	.92
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.04	0.07	[-0.18, 0.11]	-0.02	.62
		Condition ⇒ Study Quality	-0.05	0.10	[-0.25, 0.15]	-0.02	.62
		Study Quality ⇒ Credibility impressions	0.72	0.03	[0.66, 0.78]	0.76	< .001
	Direct	Condition ⇒ Credibility impressions	-0.05	0.06	[-0.16, 0.06]	-0.02	.42
	Total	Condition ⇒ Credibility impressions	-0.10	0.09	[-0.28, 0.08]	-0.05	.26
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.23	0.12	[-0.47, 0.00]	-0.11	.056
		Condition ⇒ Study Quality	-0.29	0.15	[-0.58, 0.00]	-0.14	.054
		Study Quality ⇒ Credibility impressions	0.81	0.04	[0.74, 0.88]	0.80	< .001
	Direct	Condition ⇒ Credibility impressions	0.02	0.09	[-0.15, 0.19]	0.01	.80
	Total	Condition ⇒ Credibility impressions	-0.19	0.13	[-0.44, 0.06]	-0.10	.14

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

***Influence of prior efficacy beliefs*.**

Next, I conducted a similar suite of analyses using the measure of prior efficacy beliefs instead of the prior support measure.

Study 1a. The results of this regression analysis were nearly identical to those using prior support. There was not a significant main effect of condition, $F(1, 462) = 1.90$, $p = .17$, $\eta_p^2 = .00$, but there was a main effect of prior efficacy beliefs, $F(1, 462) = 4.53$, $p = .034$, $\eta_p^2 = .01$, qualified by a significant interaction, $F(1, 462) = 13.50$, $p < .001$, $\eta_p^2 = .03$. An analysis of the estimated marginal means indicated that participants with below-average efficacy beliefs formed significantly more positive credibility impressions in the blinded condition ($M = 4.52$, 95% CI [4.34, 4.71]) relative to the unblinded condition ($M = 4.06$, 95% CI [3.89, 4.24]). Participants with average prior support also had more positive credibility impressions in the blinded condition ($M = 4.66$, 95% CI [4.53, 4.78]) than in the unblinded condition ($M = 4.53$, 95% CI [4.40, 4.66]), although the difference between conditions was not significant for these participants. Those with above-average support for capital punishment had slightly more *negative* credibility impressions in the blinded condition ($M = 4.79$, 95% CI [4.62, 4.96]) than in the unblinded condition ($M = 5.00$, 95% CI [4.81, 5.19]), but the difference between conditions was not significant for these participants.

The moderated mediation analysis using prior efficacy beliefs, presented in Table 1.6, also yielded highly similar results to the model using prior support for Study 1a. That is, there was a nonsignificant total, indirect, and direct effect of condition on credibility impressions for participants with stronger prior efficacy beliefs (above-average, $M + 1SD$). In contrast, there were significant indirect and direct effects for participants with average prior efficacy beliefs, and there were significant total, indirect, and direct effects of condition on credibility

impressions for participants with stronger prior beliefs in the inefficacy of capital punishment (below-average, M - 1SD). Most critically, for those with average and below-average prior efficacy beliefs, being in the unblinded condition led participants to provide more negative study quality evaluations than their attitudinally-equivalent counterparts in the blinded condition, and this further drove these participants to form more negative credibility impressions. While being in the unblinded condition also had a significant direct effect of making these participants' credibility impressions more positive, the indirect effects of being in the unblinded condition were substantially larger than the direct effects, making the total effect of being in the unblinded condition negative for both groups (though only significantly negative for participants with below-average prior efficacy beliefs). Thus, knowing the results of the study when making evaluations caused participants for whom the presented results were politically-unfriendly to have more negative credibility impressions of the study, and it did this by allowing these participants to provide more negatively biased study quality evaluations than their blinded counterparts.

Study 1b. Unlike in Study 1a, there were no significant effects of condition or prior efficacy beliefs on credibility impressions in Study 1b. The linear regression model indicated that neither of the main effects nor the interaction effect were significant (ps $\geq$ .34), and the moderated mediation model revealed no significant indirect, direct, nor total effects of condition on credibility impressions across levels of prior efficacy beliefs (full model details are presented in Appendix A).

In total, participants who were presented with results that conflicted with their prior beliefs formed more negative credibility impressions of the presented evidence when in the unblinded condition in Study 1a, and this effect was explained by the negatively biased study

quality evaluations these participants provided compared to their blinded counterparts. This pattern of results did not replicate in Study 1b.

Table 1.6. *Moderated mediation estimates of condition predicting credibility impressions by prior efficacy beliefs for Study 1a.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.67	0.11	[-0.89, -0.45]	-0.33	< .001
		Condition ⇒ Study Quality	-0.97	0.14	[-1.24, -0.69]	-0.45	< .001
		Study Quality ⇒ Credibility impressions	0.69	0.04	[0.61, 0.78]	0.74	< .001
	Direct	Condition ⇒ Credibility impressions	0.22	0.10	[0.03, 0.41]	0.11	.026
	Total	Condition ⇒ Credibility impressions	-0.46	0.13	[-0.71, -0.21]	-0.22	< .001
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.28	0.07	[-0.42, -0.14]	-0.13	< .001
		Condition ⇒ Study Quality	-0.38	0.09	[-0.56, -0.20]	-0.18	< .001
		Study Quality ⇒ Credibility impressions	0.73	0.03	[0.67, 0.80]	0.76	< .001
	Direct	Condition ⇒ Credibility impressions	0.13	0.06	[0.00, 0.23]	0.06	.041
	Total	Condition ⇒ Credibility impressions	-0.13	0.09	[-0.30, 0.05]	-0.06	.17
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.16	0.10	[-0.04, 0.36]	0.08	.10
		Condition ⇒ Study Quality	0.21	0.13	[-0.05, 0.46]	0.10	.10
		Study Quality ⇒ Credibility impressions	0.78	0.04	[0.69, 0.86]	0.78	< .001
	Direct	Condition ⇒ Credibility impressions	0.04	0.10	[-0.14, 0.23]	0.02	.64
	Total	Condition ⇒ Credibility impressions	0.21	0.13	[-0.04, 0.46]	0.10	.10

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Updating support and efficacy beliefs

Finally, I conducted a series of analyses to assess whether participants updated their actual support or efficacy beliefs. For these analyses, the primary outcome variables were the difference between participants' final beliefs (measured after participants completed the credibility impressions items) and their prior beliefs.[3] Thus, I constructed linear regression models with condition (dummy-coded, 0 = blinded), the relevant prior belief measure (mean-centered), and their interaction as predictors of the respective beliefs change outcomes. These analyses were conducted to assess whether the blinding manipulation influenced belief updating and whether that effect varied for participants with different prior beliefs. Moreover, I conducted moderated mediation analyses with each outcome variable to assess whether any differences in the impact of the blinding manipulation on belief updating between levels of prior beliefs could be explained by study quality evaluations.

Updating support beliefs.

Study 1a. In the regression model for Study 1a, the intercept (b = 0.13, SE = 0.06, 95% CI [0.01, 0.24]) indicated that participants slightly increased their support for capital punishment in the blinded condition. There was not a main effect of condition, $F(1, 462) = 0.25$, $p = .62$, $\eta_p^2 = .00$, but there was a large main effect or prior support, $F(1, 462) = 9.39$, $p = .002$, $\eta_p^2 = .02$, such that greater prior support for capital punishment was predictive of less attitude change, b = -0.09, SE = 0.03, 95% CI [-0.14, -0.03], β = -0.19. In other words, participants for whom the presented results were politically-friendly changed their attitudes less than those for whom the results were less politically-friendly. The interaction was not significant, $F(1, 462) = 0.00$, p =

[3] Analyses using the perceived belief change measures are described in more detail in Appendix A. The direction of effects was not consistent with the measures of actual belief change, suggesting that participants were not particularly accurate in assessing their own belief change.

.96, $\eta_p^2 = .00$, indicating that the lack of effect of condition on change in support beliefs did not vary as a function of participants' initial support beliefs.

However, the moderated mediation model showed that, while the amount of belief updating that occurred did not vary by condition or participants' prior support, the processes by which participants updated their support beliefs did significantly vary across these factors. The estimates of the moderated mediation model are presented in Table 1.7. Participants who entered the study more supportive of capital punishment (above-average support) did not have any significant direct, indirect, or total effects of condition on belief updating. The blinding manipulation did not significantly alter how these participants updated their support beliefs. Alternatively, the blinding manipulation had countervailing indirect and direct effects on the belief updating of participants with average and below-average prior support. The indirect effects indicated that participants who came into the study less supportive of capital punishment provided significantly more negative quality evaluations in the unblinded condition, which lessened the amount of belief change that occurred for these participants. Yet making unblinded evaluations also had an offsetting direct effect on belief updating, such that being in the unblinded condition modestly increased the amount of belief updating that occurred for participants with average and below-average prior support. Despite being nonsignificant, the direct effects of condition counteracted the significant indirect effects of condition, resulting in nonsignificant total effects of condition on change in support beliefs for these participants. Thus, blinding the quality evaluations of participants who viewed politically-unfriendly results influenced them to change their beliefs in the direction of the presented evidence, yet countervailing effects of the blinding manipulation ultimately reduced the how much these participants updated their beliefs.

Study 1b. The analyses of change in support beliefs for Study 1b yielded similar results. The intercept (b = -1.16, SE = 0.09, 95% CI [-1.34, -0.98]) indicated that participants generally decreased their support for trigger warnings in the blinded condition. As in Study 1a, there was not a main effect of condition, $F(1, 459) = 0.58$, $p = .45$, $\eta_p^2 = .00$, yet there was a main effect or prior support, $F(1, 459) = 38.18$, $p < .001$, $\eta_p^2 = .08$, such that prior support for trigger warnings was predictive of more attitude change, b = -0.38, SE = 0.06, 95% CI [-0.51, -0.26], β = -0.40. In other words, participants who were presented with politically-unfriendly results changed their support beliefs more than participants for whom the results were more politically-friendly. The interaction was not significant, $F(1, 459) = 1.03$, $p = .31$, $\eta_p^2 = .00$, indicating that the lack of effect of condition on change in support beliefs did not vary as a function of participants' initial support beliefs.

The moderated mediation model of Study 1b, presented in Table 1.8, also showed a comparable pattern of results as Study 1a. Participants with above-average prior support for trigger warnings—for whom the presented results were politically-unfriendly—had lower study quality evaluations in the unblinded condition than in the blinded condition. Those deflated study quality evaluations subsequently reduced the how much unblinded participants with above-average prior support updated their support beliefs. However, while both components of this indirect effect were statistically significant, the overall indirect effect was not significant, suggesting a lack of statistical power. Thus, Study 1b produced weak evidence in line with the results of Study 1a: the blinding manipulation influenced how participants who received politically-unfriendly information updated their support beliefs, but these condition effects did not ultimately result in differences in belief change for these participants.

Table 1.7. *Moderated mediation estimates of condition predicting change in support beliefs by prior support beliefs for Study 1a.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	-0.19	0.07	[-0.34, -0.05]	-0.11	.008
		Condition ⇒ Study Quality	-0.93	0.13	[-1.19, -0.66]	-0.43	< .001
		Study Quality ⇒ Change in Support	0.21	0.07	[0.07, 0.35]	0.25	.004
	Direct	Condition ⇒ Change in Support	0.15	0.14	[-0.11, 0.43]	0.09	.26
	Total	Condition ⇒ Change in Support	-0.04	0.11	[-0.26, 0.19]	-0.02	.75
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	-0.06	0.02	[-0.11, -0.01]	-0.03	.01
		Condition ⇒ Study Quality	-0.38	0.09	[-0.57, -0.20]	-0.18	< .001
		Study Quality ⇒ Change in Support	0.16	0.05	[0.06, 0.25]	0.20	.001
	Direct	Condition ⇒ Change in Support	0.05	0.09	[-0.13, 0.22]	0.03	.62
	Total	Condition ⇒ Change in Support	-0.04	0.08	[-0.20, 0.12]	-0.02	.61
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	0.02	0.02	[-0.02, 0.06]	0.01	.36
		Condition ⇒ Study Quality	0.16	0.13	[-0.09, 0.42]	0.08	.22
		Study Quality ⇒ Change in Support	0.11	0.06	[0.00, 0.23]	0.14	.062
	Direct	Condition ⇒ Change in Support	-0.06	0.11	[-0.27, 0.14]	-0.04	.55
	Total	Condition ⇒ Change in Support	-0.04	0.11	[-0.27, 0.18]	-0.03	.70

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 1.8. *Moderated mediation estimates of condition predicting change in support beliefs by prior support beliefs for Study 1b.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	-0.02	0.03	[-0.08, 0.04]	-0.01	.42
		Condition ⇒ Study Quality	0.19	0.13	[-0.07, 0.44]	0.09	.15
		Study Quality ⇒ Change in Support	-0.13	0.10	[-0.31, 0.07]	-0.09	.20
	Direct	Condition ⇒ Change in Support	-0.04	0.19	[-0.43, 0.31]	-0.01	.83
	Total	Condition ⇒ Change in Support	-0.03	0.18	[-0.39, 0.33]	-0.01	.85
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	0.01	0.03	[-0.04, 0.07]	0.00	.62
		Condition ⇒ Study Quality	-0.05	0.10	[-0.24, 0.14]	-0.02	.60
		Study Quality ⇒ Change in Support	-0.27	0.07	[-0.40, -0.13]	-0.19	< .001
	Direct	Condition ⇒ Change in Support	0.05	0.13	[-0.20, 0.30]	0.02	.67
	Total	Condition ⇒ Change in Support	0.10	0.13	[-0.16, 0.35]	0.03	.45
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	0.12	0.07	[-0.02, 0.25]	0.04	.08
		Condition ⇒ Study Quality	-0.29	0.15	[-0.57, -0.00]	-0.14	.049
		Study Quality ⇒ Change in Support	-0.41	0.10	[-0.60, -0.20]	-0.28	< .001
	Direct	Condition ⇒ Change in Support	0.15	0.19	[-0.23, 0.53]	0.05	.45
	Total	Condition ⇒ Change in Support	0.23	0.18	[-0.13, 0.59]	0.08	.21

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

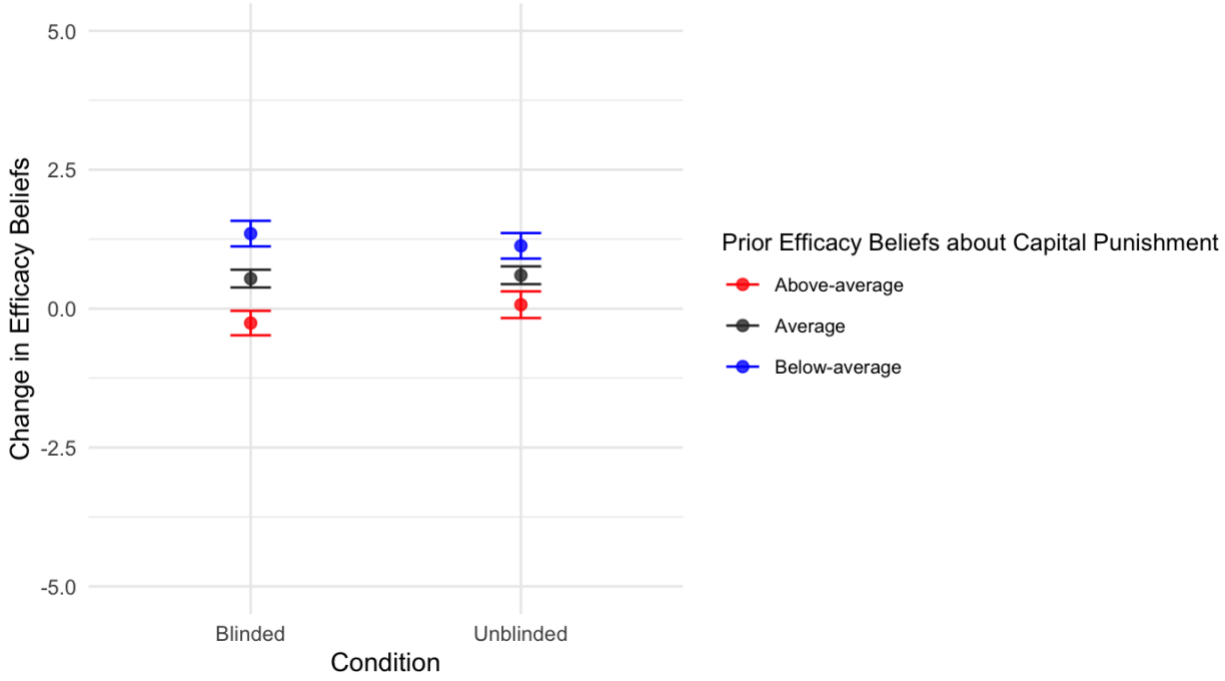
Updating efficacy beliefs

Study 1a. In the Study 1a regression model, the intercept (b = 0.54, SE = 0.08, 95% CI [0.38, 0.70]) indicated that participants generally increased their belief in the effectiveness of capital punishment in the blinded condition. There was not a main effect of condition, $F(1, 462) = 0.22$, $p = .64$, $\eta_p^2 = .00$, but there was a large main effect or prior support, $F(1, 462) = 103.55$, $p < .001$, $\eta_p^2 = .18$, such that greater prior beliefs in the efficacy of capital punishment were associated with less attitude change, b = -0.09, SE = 0.03, 95% CI [-0.14, -0.03], β = -0.19. Participants for whom the presented results more closely matched their prior beliefs changed their efficacy beliefs less than those for whom the results were less expected. However, the interaction term was significant in this analysis, $F(1, 462) = 0.00$, $p = .96$, $\eta_p^2 = .00$. This interaction showed that participants with stronger prior efficacy beliefs changed their efficacy beliefs significantly less in the blinded condition, b = -0.43, SE = 0.04, 95% CI [-0.52, -0.35], β = -0.57, than in the unblinded condition, b = -0.28, SE = 0.05, 95% CI [-0.37, -0.19], β = -0.37.

Moreover, the estimated marginal means are illustrated in Figure 1.10. Participants with average and below-average prior efficacy beliefs updated their beliefs in the direction of the presented evidence, and this did not differ across conditions. On the other hand, those with above-average prior efficacy beliefs, for whom the results were politically-friendly, came to believe that capital punishment is significantly *less* effective than they originally believed in the blinded condition, but these participants did not change their beliefs in the unblinded condition.

Figure 1.10. Average Change in Efficacy Beliefs by Condition and Prior Efficacy Beliefs in Study 1a

Error bars represent 95% CIs



The results of the moderated mediation, presented in Table 1.9, indicated that the processes by which participants updated their efficacy beliefs significantly varied as a function of condition and participants' prior efficacy beliefs. Participants with average and below-average prior support rated the study as being of lower quality in the unblinded condition, which decreased the amount of belief change that occurred for these participants. Yet there were also significant, countervailing direct effects of condition on change in efficacy beliefs for these participants, such that being in the unblinded condition modestly increased how much these participants came to believe that capital punishment is effective. The direct effects of condition cancelled out the negative indirect effects of condition, resulting in nonsignificant total effects of condition on change in efficacy beliefs for participants with average or below-average prior efficacy beliefs. In contrast, there was a significant total effect of condition on change in efficacy beliefs for participants with above-average prior efficacy beliefs, such that being in the unblinded

condition increased the amount that these participants came to believe that capital punishment is more effective than they originally thought.

Ultimately, the blinding manipulation affected *how* participants changed their beliefs about the efficacy of capital punishment in Study 1a. Participants with average and below-average prior efficacy beliefs evaluated the study more positively in the blinded condition, yet they also relied more strongly on their prior beliefs (and, consequently, less on the new evidence) when forming their final efficacy beliefs. These countervailing effects resulted in nonsignificant differences in belief change across conditions for these participants. Blinding also reduced the *amount* of belief updating that occurred for participants with above-average prior efficacy beliefs by reducing their reliance on their prior beliefs when forming their final beliefs.

Study 1b. The models predicting change in efficacy beliefs Study 1b did not show the same moderation effects that were observed in Study 1a. The intercept of the regression model (b = -1.31, SE = 0.09, 95% CI [-1.49, -1.14]) indicated that participants in the blinded condition generally came to believe that trigger warnings are less effective than they originally thought. There was not a main effect of condition, $F(1, 459) = 0.20$, $p = .66$, $\eta_p^2 = .00$, but there was a main effect or prior support, $F(1, 459) = 89.80$, $p < .001$, $\eta_p^2 = .16$, such that greater prior belief in the efficacy of trigger warnings was predictive of greater belief change in the direction of the presented evidence, b = -0.55, SE = 0.06, 95% CI [-0.66, -0.43], β = -0.54. However, the interaction was not significant, $F(1, 459) = 0.10$, $p = .75$, $\eta_p^2 = .00$, unlike the same model in Study 1a, indicating that prior efficacy beliefs were equivalently predictive of attitude change across conditions. Moreover, the results of the moderated mediation analysis, provided in Table 1.10, showed nonsignificant indirect, direct, or total effects of condition on change in efficacy beliefs across levels of prior efficacy beliefs. Thus, the more the presented results ran counter to

what participants originally believed, the more likely they were to update their efficacy beliefs in Study 1b, and the process by which participants updated their efficacy beliefs did not differ across conditions.

Table 1.9. *Moderated mediation estimates of condition predicting change in efficacy beliefs by prior efficacy beliefs for Study 1a.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	-0.55	0.11	[-0.77, -0.35]	-0.19	< .001
		Condition ⇒ Study Quality	-0.97	0.14	[-1.25, -0.71]	-0.45	< .001
		Study Quality ⇒ Change in Efficacy	0.57	0.07	[0.44, 0.70]	0.42	< .001
	Direct	Condition ⇒ Change in Efficacy	0.31	0.17	[-0.02, 0.64]	0.11	.063
	Total	Condition ⇒ Change in Efficacy	-0.23	0.16	[-0.55, 0.10]	-0.08	.17
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	-0.15	0.04	[-0.24, -0.07]	-0.05	< .001
		Condition ⇒ Study Quality	-0.38	0.09	[-0.56, -0.21]	-0.18	< .001
		Study Quality ⇒ Change in Efficacy	0.41	0.05	[0.30, 0.51]	0.31	< .001
	Direct	Condition ⇒ Change in Efficacy	0.30	0.12	[0.07, 0.53]	0.11	.009
	Total	Condition ⇒ Change in Efficacy	0.05	0.12	[-0.17, 0.28]	0.02	.64
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	0.05	0.04	[-0.02, 0.13]	0.02	.18
		Condition ⇒ Study Quality	0.21	0.13	[-0.03, 0.46]	0.10	.097
		Study Quality ⇒ Change in Efficacy	0.24	0.08	[0.09, 0.39]	0.19	.002
	Direct	Condition ⇒ Change in Efficacy	0.29	0.16	[-0.03, 0.60]	0.10	.074
	Total	Condition ⇒ Change in Efficacy	0.33	0.16	[0.01, 0.66]	0.12	.041

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 1.10. *Moderated mediation estimates of condition predicting change in efficacy beliefs by prior efficacy beliefs for Study 1b.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	-0.04	0.04	[-0.13, 0.05]	-0.01	.37
		Condition ⇒ Study Quality	0.12	0.15	[-0.17, 0.41]	0.06	.39
		Study Quality ⇒ Change in Efficacy	-0.32	0.08	[-0.47, -0.18]	-0.20	< .001
	Direct	Condition ⇒ Change in Efficacy	0.04	0.15	[-0.25, 0.36]	0.01	.77
	Total	Condition ⇒ Change in Efficacy	0.02	0.18	[-0.34, 0.37]	0.01	.93
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	0.01	0.04	[-0.06, 0.08]	0.00	.75
		Condition ⇒ Study Quality	-0.03	0.10	[-0.22, 0.16]	-0.01	.75
		Study Quality ⇒ Change in Efficacy	-0.36	0.06	[-0.49, -0.24]	-0.23	< .001
	Direct	Condition ⇒ Change in Efficacy	0.04	0.12	[-0.20, 0.29]	0.01	.75
	Total	Condition ⇒ Change in Efficacy	0.06	0.13	[-0.19, 0.31]	0.02	.65
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	0.08	0.07	[-0.06, 0.21]	0.02	.26
		Condition ⇒ Study Quality	-0.19	0.15	[-0.47, 0.11]	-0.09	.21
		Study Quality ⇒ Change in Efficacy	-0.41	0.11	[-0.62, -0.20]	-0.26	< .001
	Direct	Condition ⇒ Change in Efficacy	0.03	0.21	[-0.39, 0.45]	0.01	.87
	Total	Condition ⇒ Change in Efficacy	0.10	0.18	[-0.26, 0.46]	0.03	.59

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Discussion

In Studies 1a and 1b, participants' prior beliefs significantly influenced their evaluations of a study's methodological quality when they were privy to its results. This was not the case for participants who offered blinded evaluations of quality, made before they knew the study's results. These findings indicated that, on average, unblinded participants provided biased evaluations of quality, for they were influenced by their prior beliefs when making those evaluations significantly more than unblinded individuals with equivalent prior beliefs. This pattern of results was clearest in Study 1a, where participants for whom the presented results were most politically-unfriendly provided significantly lower quality evaluations when they made unblinded evaluations. Some similar patterns were observed in Study 1b, although the effects were generally smaller and did not always reach statistical significance. These studies provided initial evidence against strong rationalist accounts of partisan reasoning and provided initial support of my hypothesis related to Research Question 1. Partisans evaluated scientific evidence in a biased manner when they knew how the presented evidence aligned with their prior beliefs.

In relation to Research Question 2, I hypothesized that both prior support and prior efficacy beliefs would bias participants' unblinded quality evaluations in each study. That hypothesis was fully supported only in Study 1a. In that study, participants' prior support for capital punishment and their prior beliefs about the efficacy of capital punishment biased their study quality evaluations. It appeared that prior efficacy beliefs were the driver of this bias, for the interaction between prior support and condition did not predict quality evaluations when the interaction between prior efficacy beliefs and condition was in the same model. This suggested that the observed biases observed in Study 1a maybe more cognitively-driven. Yet in Study 1b,

only prior support for trigger warnings exerted a biasing influence on unblinded quality evaluations, which suggests a more affectively-driven account of partisan bias. However, Study 1b had less precise estimates across models, and I lacked adequate statistical power for many of the analyses including interaction effects in Study 1a, including the moderated mediation models. Thus, I aimed to recruit larger samples in my subsequent studies to more precisely estimate the relative biasing influences of these beliefs on quality evaluations and subsequent belief-updating.

These studies also provided evidence pertinent to Research Question 3, showing that unblinded participants' biased evaluations had downstream influences on their credibility impressions and belief updating. Participants who viewed politically-unfriendly information in Study 1a formed significantly more positive credibility impressions (e.g., found the evidence to be more believable) when they evaluated the study's methodological quality without knowing its results. These participants also changed their beliefs about the efficacy of capital punishment more when they provided blinded evaluations, although this difference was not statistically significant. Moderated mediation analyses showed that these effects could be explained by the indirect effect that blinded participants' elevated quality evaluations had on these outcomes, yet the blinding manipulation also directly deflated these participants' credibility impressions and belief updating. Similar patterns were observed in Study 1b, although the indirect effects were consistently smaller and did not reach statistical significance. It is not clear why the blinding manipulation induced negative direct effects on credibility impressions and belief updating for participants who viewed politically-unfriendly results. Perhaps separating the information about the study methods and results felt more artificial when the results were politically-unfriendly, which caused these participants to distrust the presented information more than their unblinded counterparts. Whether or not that conjecture helps explain the effect, the manipulation caused

these participants—but not those for whom the results were politically-friendly—to perceive the presented information as having lower evidentiary value. In the General Discussion, I will return to discussing potential explanations for these effects. Nevertheless, in aggregate, the moderated mediation models suggested that the blinding manipulation may make people slightly more responsive to politically-unfriendly results by prohibiting them from making biased evaluations of its methodological merits.

STUDY 2

In Study 2, I recruited twice as large of a sample to replicate and extend the findings of Studies 1a and 1b. I used the same methods and materials as Study 1a, and the preregistered hypotheses were identical to those of the previous studies as well. Namely, I hypothesized that prior beliefs would significantly predict quality evaluations only in the unblinded condition (addressing Research Question 1) and that both prior support and prior efficacy beliefs would significantly bias unblinded participants' quality evaluations (addressing Research Question 2). Additionally, I preregistered analyses to examine how the blinding manipulation and participants' prior beliefs influenced their credibility impressions and belief updating (addressing Research Question 3).

I also preregistered several secondary analyses related to Research Question 2. First, I measured participants emotions just before they completed their quality evaluations, and I planned analyses to assess whether positive and/or negative affective reactions could account for the biasing influence of prior beliefs on quality evaluations in the unblinded condition. This would provide stronger evidence of affectively-driven biases than I had previously demonstrated. Similarly, I also measured participants' moral conviction about capital punishment to determine whether more morally convicted participants exhibited stronger biases. Lastly, to examine the role that reflective thinking may have on partisan bias, I also collected participants' responses to an analytic thinking task. There are competing accounts of whether increased reflectiveness is associated with stronger (Kahan et al., 2017), weaker (Pennycook & Rand, 2019; Tappin et al., 2020a) or both stronger and weaker (Batailler et al., 2022) partisan biases, so I examined the how analytic thinking was related to partisans' evaluations to help refine these theoretical claims.

Method

Participants

A power analysis indicated that, to have 0.80 power to detect small ($f^2 = 0.02$) three-way interactions, I would need to recruit at least 725 participants. Anticipating that some participants would not pass our preregistered inclusion criteria (three English comprehension questions, a manipulation check, and completing the survey in less than three minutes), I aimed to recruit 1000 participants through Prolific.

Ultimately, 1007 participants were recruited in June of 2021. The participants ranged in age from 18-85 years ($M = 38.52$, $SD = 13.54$), ranged in yearly household income from less than $5,000 to over $175,000 (*Mdn* = $50,000 - $59,999) and were mainly White (72%), female (52%), and college-educated (37%).

Procedure and Measures

After consenting, completing a captcha, and responding to three English comprehension questions, participants were asked the same prior support, prior efficacy beliefs, and moral conviction questions as were used in Study 1a. Participants were then randomly assigned to one of two experimental conditions (Condition: blinded or unblinded) in a two-cell between-subjects design. All participants read a brief introduction and methods description of the focal study about capital punishment, as was presented in Study 1a. Participants in the unblinded condition were only required to stay on the page containing the initial study description for 30 seconds before they could proceed to the next page. Those in the blinded condition were required to stay on the page for 30 seconds but were also presented the 20-item PANAS (Watson et al., 1988) and the study quality items immediately after the methods description. The wordings of the study quality items for this study were identical to the items used in Study 1a.

On the next page, all participants were presented with the full write-up of the study, which included the introduction and methods descriptions and a brief description of the results. As in Study 1a, all participants read that the study found that capital punishment was an effective deterrent of violent crime. All participants were required to stay on this page for at least 30 seconds before proceeding. Participants in the unblinded conditions completed the PANAS and the study quality measures on this page, while participants in the blinded conditions were not presented any items before proceeding.

All participants were then presented with the manipulation check and the credibility impressions items from Study 1a. On the following page, participants were asked to complete the second capital punishment support and efficacy belief items, which were used as the second time point for measures of actual belief change. On the subsequent page, participants were asked the same reported belief change measures used in Study 1a for reported change in support and efficacy beliefs, respectively. Participants were then provided with an optional open response question to share any additional information about their views of the study they read or the issue of capital punishment more broadly.

Finally, participants completed the Cognitive Reflection Task 1 and 2 (CRT; Frederick, 2005; Thomson & Oppenheimer, 2016) and then completed the same demographic information about their age, sex, ethnicity, income, education, political orientation (social and economic, separately), and political party affiliation as in Study 1a. Participants were given the opportunity to share any thoughts or feelings about the study in an open-response question before being presented with the debriefing and completing the survey.

Results

Following my preregistration, I excluded participants who did not pass our English comprehension checks, failed our manipulation check (responding "*Capital punishment is not effective at all*"), or finished the survey in less than three minutes. This resulted in a sample of 912 participants for the confirmatory analyses (including the full sample did not substantively alter the results). The distribution of participants across the two experimental conditions was roughly equivalent ($n_{blinded}$ = 455, $n_{unblinded}$ = 457).

On average, participants were slightly opposed to capital punishment (*M* = 3.57, *SD* = 1.99) and thought that capital punishment is somewhat effective at deterring violent crime (*M* = 3.24, *SD* = 1.90). As in Study 1a, these items were strongly correlated (*r* = 0.76). The measure of moral conviction had suitable internal reliability (Cronbach's α = 0.76), and participants felt morally convicted about the issue on average (*M* = 4.86, *SD* = 1.46). Participants also reported being slightly liberal on social (*M* = 5.00, *SD* = 1.88) and economic (*M* = 4.71, *SD* = 1.93) issues and leaned Democrat (*M* = 4.90, *SD* = 1.68). As in the previous studies, the social and economic political orientation measures were combined into a composite political orientation measure (Cronbach's α = 0.92). Also as in the previous studies, indices of skewness and kurtosis indicated that the distributions of responses were sufficiently normal for parametric analyses (Field et al., 2012).

Study quality evaluations

The study quality items had high internal reliability in Study 2 (Cronbach's α = 0.90) and were thus averaged into a composite measure[4]. On average, participants rated the study as being

[4] As in Studies 1a and 1b, the internal reliability of the quality items in the blinded condition (Cronbach's α = 0.88) did not meaningfully differ from the reliability in the unblinded condition (Cronbach's α = 0.91).

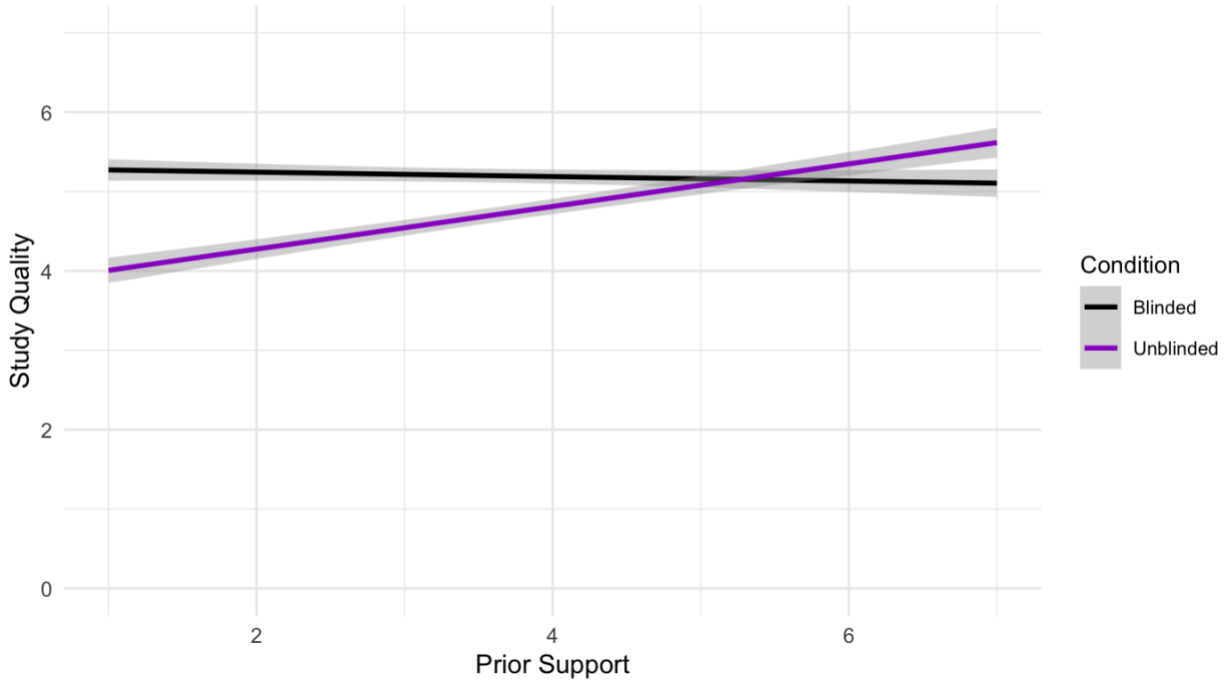
of decent quality (M = 4.96, SD = 1.07). To test Research Questions 1 and 2, I conducted a series of linear regression analyses identical to those I ran in Studies 1a and 1b.

Influence of prior support beliefs

First, I constructed a model predicting study quality evaluations from condition (dummy-coded, 0 = blinded), prior support for capital punishment (mean-centered), and their interaction. Omnibus tests showed a significant overall model, $F(3, 908) = 62.81$, $p < .001$, $\eta_p^2 = .17$, and a main effect of condition $F(1, 908) = 60.45$, $p < .001$, $\eta_p^2 = .06$, but not of prior support, $F(1, 908) = 1.49$, $p = .22$, $\eta_p^2 = .00$. However, this was qualified by a highly significant interaction between prior support and condition, $F(1, 908) = 82.30$, $p < .001$, $\eta_p^2 = .08$, illustrated in Figure 2.1. Simple effects analyses showed that, as hypothesized, prior support for capital punishment was predictive of more positive study quality evaluations in the unblinded condition, b = 0.27, SE = 0.02, $p < .001$, 95% CI [0.22, 0.31], β = 0.50, but not in the blinded condition, b = -0.03, SE = 0.02, p = .22, 95% CI [-0.07, 0.02], β = -0.05. While blinded participants' prior support was unrelated to their quality evaluations, participants who knew the study's results when submitting their evaluations were significantly influenced by their prior support when evaluating the study. In other words, participants in the unblinded condition significantly deviated from the accuracy-motivated baselines of quality evaluations provided by participants in the blinded condition, revealing a significant directional bias in their evaluations.

Figure 2.1. Average Study Quality Evaluations by Prior Support and Condition in Study 2

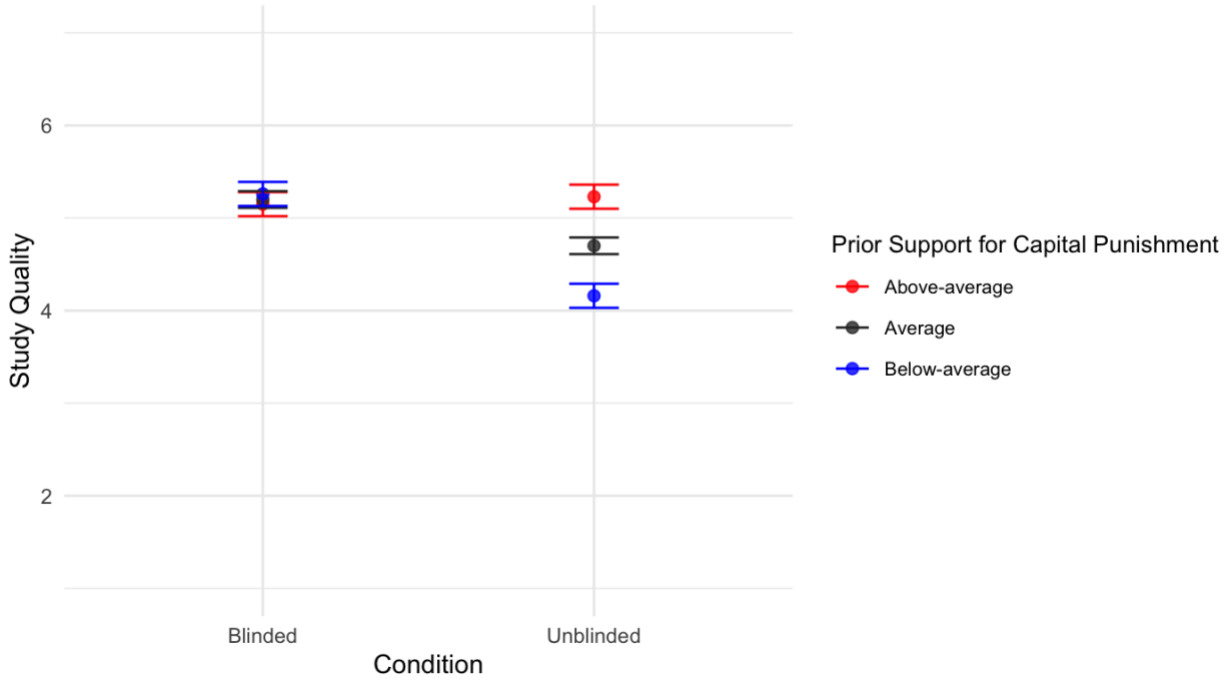
Error bands represent standard errors



Additionally, I compared the estimated marginal means of participants with below-average (M – 1SD), average, and above-average (M + 1SD) prior support across conditions. Figure 2.2 presents the estimated marginal means and corresponding 95% confidence intervals. While there were not significant differences in quality evaluations between conditions for those with above-average prior support, there were significant differences (as indicated by nonoverlapping 95% CIs) across conditions for those with either average or below-average prior support. That is, participants for whom the presented results were politically-unfriendly provided negatively biased quality evaluations in the unblinded condition relative to their attitudinally-similar counterparts in the blinded condition, but participants for whom the results aligned with their prior support did not demonstrate a significant bias in their unblinded evaluations.

Figure 2.2 Average Study Quality Evaluations by Condition and Prior Support in Study 2

Error bars represent 95% CIs



Influence of prior efficacy beliefs

Next, I constructed a model predicting study quality evaluations from condition (dummy-coded, 0 = blinded/conservative-friendly), prior efficacy beliefs about capital punishment (mean-centered), and their interaction. Omnibus tests showed a significant overall model, $F(3, 908) = 75.70$, $p < .001$, $\eta_p^2 = .20$, and a main effect of condition $F(1, 908) = 59.09$, $p < .001$, $\eta_p^2 = .06$, but not of prior efficacy beliefs, $F(1, 908) = 0.13$, $p = .72$, $\eta_p^2 = .00$. This was qualified by a highly significant interaction between condition and prior efficacy beliefs, $F(1, 908) = 93.69$, $p < .001$, $\eta_p^2 = .09$. As predicted, simple effects analyses showed that prior belief in the efficacy of capital punishment was predictive of more positive quality evaluations in the unblinded condition, $b = 0.32$, $SE = 0.02$, $p < .001$, 95% CI [0.27, 0.37], $\beta = 0.56$, but not in the blinded condition, $b = -0.01$, $SE = 0.02$, $p = .72$, 95% CI [-0.05, 0.04], $\beta = -0.01$. Blinded participants'

prior efficacy beliefs unrelated to their quality evaluations, but unblinded participants evaluations were significantly biased by their prior efficacy beliefs when evaluating the study.

In comparing the estimated marginal means of participants with below-average, average, and above-average prior efficacy beliefs across conditions, I found similar results as in the analysis of prior support. While there were not significant differences in quality evaluations between conditions for those high in prior efficacy beliefs, there were significant differences (as indicated by nonoverlapping 95% CIs) across conditions for those with either average or low prior efficacy beliefs. Table 2.1 presents these estimated marginal means and corresponding 95% confidence intervals. In sum, aligning with my hypotheses regarding Research Question 1, prior efficacy beliefs exerted a biasing influence on unblinded participants' quality evaluations, particularly for participants for whom the presented results were more politically-unfriendly.

Table 2.1. *Estimated Marginal Means of Study Quality Evaluations in Study 2 by Condition and Prior Efficacy Beliefs*

	Condition	
Prior Efficacy Beliefs Group	Blinded	Unblinded
Below-average (Less effective)	5.22 [5.10, 5.34]	**4.11 [3.98, 4.24]**
Average	5.20 [5.11, 5.29]	**4.71 [4.62, 4.80]**
Above-average (More effective)	5.19 [5.06, 5.31]	5.32 [5.19, 5.44]

Note: Values in brackets are 95% confidence intervals. Bolded values indicate nonoverlapping confidence intervals from the Blinded condition.

Relative influence of prior support and prior efficacy beliefs

To further address Research Question 2, I assessed the relative influence of prior support and prior efficacy beliefs on quality evaluations. I constructed a regression model using the condition variable, both prior belief measures, and the two-way interactions between condition and each prior belief variable. The key terms in this model are the interaction terms, which

indicate the influence that each prior belief measure had on quality evaluation when accounting for the shared variance between the two measures.

The overall model was significant, $F(5, 906) = 47.92$, $p < .001$, $\eta_p^2 = .21$, as was the main effect of condition, $F(1, 906) = 60.28$, $p < .001$, $\eta_p^2 = .06$. Neither the main effect of prior support, $F(1, 906) = 2.26$, $p = .13$, $\eta_p^2 = .00$, nor the main effect of prior efficacy beliefs, $F(1, 906) = 0.84$, $p = .36$, $\eta_p^2 = .00$, were significant. However, the key interactions between prior support and condition, $F(1, 906) = 9.48$, $p = .002$, $\eta_p^2 = .01$, and between prior efficacy beliefs and condition, $F(1, 906) = 15.90$, $p < .001$, $\eta_p^2 = .02$, were significant and in the same direction as the previous analyses. Above and beyond the shared variance explained by the two prior belief variables, both prior support and prior efficacy beliefs had independent influences on participants quality evaluations in the unblinded condition—but not in the blinded condition. Thus, both prior support and prior efficacy beliefs independently biased unblinded participants' quality evaluations.

However, most of the biasing influence of these prior beliefs seemed to stem from their shared influence on quality evaluations. The total variance explained by the combined model with two interaction terms ($\eta_p^2 = .21$) was only slightly larger than the variance explained by the model only including the interaction with prior support ($\eta_p^2 = .17$) or the model only including the interaction with prior efficacy beliefs ($\eta_p^2 = .20$). This suggested that the two prior belief measures were accounting for similar variance across models rather than explaining distinct variance in quality evaluations. In line with this interpretation, the interaction between condition and prior support accounted for less variance in this combined model ($\eta_p^2 = .01$) than in the model only including the prior support interaction term ($\eta_p^2 = .08$); similarly, the interaction between condition and prior efficacy beliefs also accounted for less variance in this combined

model (η_p^2 = .02) than in the model only including prior efficacy interaction term (η_p^2 = .09). Since the main effects accounted for the same amount of variance in each of the three models (η_p^2 = .06), this indicated that the remaining variance explained by the combined model (η_p^2 = .12) was due to the shared influence of prior support and prior efficacy beliefs. In summary, prior support and prior efficacy beliefs yielded distinguishable, but highly overlapping, directional biases in quality evaluations for unblinded participants.

Exploratory analysis of political orientation

I also ran an exploratory regression analysis to examine whether the results of the confirmatory analyses would replicate when using political orientation rather than prior support or prior efficacy beliefs. The full results of this analysis are presented in Appendix B. As in Studies 1a, participants' political orientation was significantly predictive of quality evaluations in the unblinded condition, but not in the blinded condition. Moreover, more liberal participants had significantly lower quality evaluations in the unblinded condition than in the blinded condition, but more conservative did not exhibit a significant bias. These provided further evidence that participants political motivations directionally biased their quality evaluations.

Mediation of study quality judgments by positive and negative affect

In relation to Research Question 2, I conducted moderated mediation analyses to assess whether positive and/or negative affect mediated the relationship between condition and study quality evaluations, and whether than relationship differed by participants' prior beliefs. Deviating from my preregistered analysis plan, which consisted of a single model including both prior support and prior efficacy beliefs, I created two separate models for each measure to better estimate the indirect effects of condition on study quality as moderated by each prior belief.

The full model estimates for these analyses are presented in Appendix B. In brief, neither positive nor negative affect accounted for substantial variance in the condition effects on study quality evaluations, and this was true across levels of participants' prior beliefs. Positive affect did account for a small proportion of the variance in the evaluations made by participants with average prior beliefs (for both the support and efficacy beliefs measures), such that being in the unblinded condition caused these participants to have lower positive affect, which in turn decreased their quality evaluations. However, these indirect effects were small, accounting for roughly 5-6% of the total effect of condition on these participants' quality evaluations. While similar patterns emerged for participants with below-average prior beliefs—for whom the presented materials were the most politically-unfriendly—these indirect effects were not significant, likely due to a lack of statistical power to detect small effects. Ultimately, these results suggest that participants' affective reactions were not the driver of the evaluative biases made by participants with average and below-average prior beliefs in the unblinded condition; the blinding manipulation did not have the requisite impact on participants' positive or negative emotions to explain the evaluative biases that were observed.

Moderation of study quality evaluations by individual difference measures

To further address Research Question 2, I constructed a series of four linear regression models to test the two-way and three-way interactions between condition, the prior beliefs measures, and the two individual difference measures. These models included the condition variable (dummy-coded, 0 = blinded), either the prior support or prior efficacy beliefs variable (mean-centered), and either moral conviction composite or CRT scores (mean-centered). These analyses were run to assess whether moral conviction or analytic thinking attenuated (or

exacerbated) the biasing influence of prior beliefs on quality evaluations in the unblinded condition.

Moderation of prior support effects

Moral conviction. Starting with moral conviction, omnibus tests showed that, in addition to the previously documented interaction between prior support and condition, $F(1, 904) = 60.42$, $p < .001$, $\eta_p^2 = .06$, there was a significant main effect of moral conviction, $F(1, 904) = 24.37$, $p < .001$, $\eta_p^2 = .03$, as well as a significant interaction between moral conviction and condition, $F(1, 904) = 6.31$, $p = .012$, $\eta_p^2 = .01$. However, there was not a significant two-way interaction between prior support and moral conviction, $F(1, 904) = 1.31$, $p = .25$, $\eta_p^2 = .00$, nor a significant three-way interaction between condition, prior support, and moral conviction, $F(1, 904) = 1.95$, $p = .16$, $\eta_p^2 = .00$.

However, simple effects analyses indicated that moral conviction may have had a small moderating effect on the relationship between prior support and quality evaluations in the unblinded condition. There was a significant difference (as indicated by nonoverlapping 95% CIs) in the influence that prior support had on unblinded participants' evaluations between those with low moral conviction, $b = 0.17$, $SE = 0.04$, $p < .001$, 95% CI [0.10, 0.24], $\beta = 0.31$, and those with high moral conviction, $b = 0.32$, $SE = 0.03$, $p < .001$, 95% CI [0.26, 0.37], $\beta = 0.59$. In other words, those who were highly morally convicted about capital punishment were significantly more influenced by their prior support when making unblinded quality evaluations than those with low moral conviction about the topic. Those with average moral conviction in the unblinded condition—which was the largest of the three moral conviction groups—did not significantly differ from either the below-average or above-average conviction groups, $b = 0.24$, $SE = 0.02$, $p < .001$, 95% CI [0.20, 0.29], $\beta = 0.45$. Thus, while there was some indication that

moral conviction exacerbated the biasing influence of prior support on quality evaluations, the lack of a significant three-way interaction indicated that I did not have adequate statistical power to reliably estimate the size of this effect.

Analytic thinking. In the analysis of analytic thinking scores, there was a significant main effect of CRT on quality evaluations, $F(1, 904) = 9.19$, $p = .003$, $\eta_p^2 = .01$, that was qualified by a significant interaction between prior support and CRT scores, $F(1, 904) = 3.89$, $p = .049$, $\eta_p^2 = .00$. While the interaction between prior support and condition remained significant, $F(1, 904) = 85.18$, $p < .001$, $\eta_p^2 = .09$, there was not an interaction between condition and CRT, $F(1, 904) = 0.50$, $p = .48$, $\eta_p^2 = .00$, nor a significant three-way interaction, $F(1, 904) = 0.81$, $p = .37$, $\eta_p^2 = .00$. Simple effects analyses did not suggest that there were any meaningful differences that I was merely underpowered to detect. In total, this analysis indicated that those higher in analytic thinking were influenced slightly less by their prior support when making quality evaluations, $b = -0.02$, $SE = 0.02$, $p = .049$, 95% CI [-0.04, -0.00], $\beta = -0.08$, but this effect did not vary across conditions.

Moderation of prior efficacy beliefs effects

Moral conviction. While there was a significant interaction between moral conviction and condition, $F(1, 904) = 9.83$, $p = .002$, $\eta_p^2 = .01$, there was not a significant three-way interaction between condition, prior efficacy beliefs, and moral conviction, $F(1, 904) = 0.94$, $p = .33$, $\eta_p^2 = .00$. However, simple effects analyses showed some indications that moral conviction may have exacerbated the influence of prior efficacy beliefs on quality evaluations in the blinded conditions, as was the case in the model using prior support. That is, the influence of prior efficacy beliefs on quality judgments was greater for those with above-average moral conviction who were in the unblinded condition, $b = 0.36$, $SE = 0.03$, $p < .001$, 95% CI [0.31, 0.42], $\beta =$

0.64, than it was for those with below-average moral conviction who were in the unblinded condition, b = 0.23, SE = 0.04, p < .001, 95% CI [0.16, 0.31], β = 0.41. In sum, there was some indication that moral conviction may have moderated the influence of prior efficacy beliefs on quality evaluations in the unblinded condition, but I did not have adequate statistical power to reliably estimate this small effect.

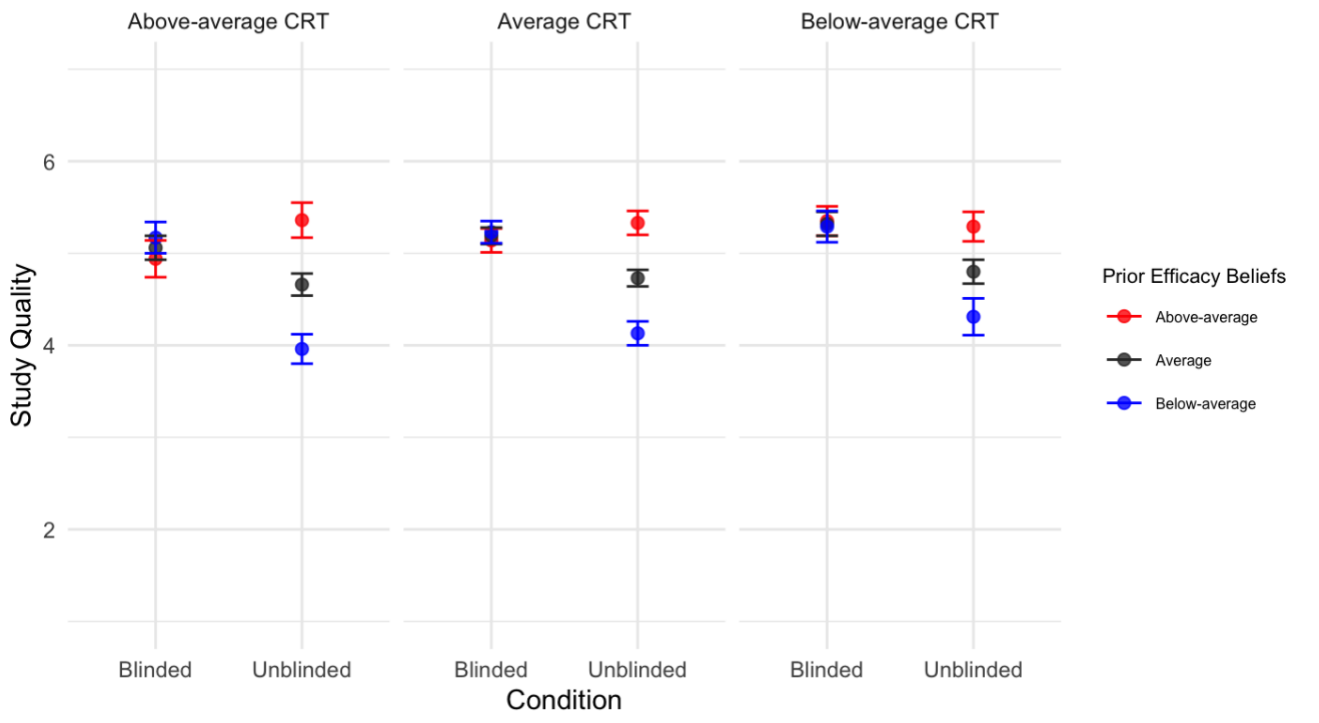
Analytic thinking. In contrast, there was a significant three-way interaction between condition, prior efficacy beliefs, and CRT scores, $F(1, 904) = 7.77$, p = .005, η_p^2 = .01. Simple effects analyses showed that the influence of prior efficacy beliefs on quality judgments was greater for those with above-average CRT scores who were in the unblinded condition, b = 0.37, SE = 0.04, p < .001, 95% CI [0.30, 0.44], β = 0.65, than it was for those with below-average CRT scores who were in the unblinded condition, b = 0.26, SE = 0.03, p < .001, 95% CI [0.19, 0.32], β = 0.45. Participants higher in analytic thinking were biased by their efficacy beliefs more than participants lower in analytic thinking.

Furthermore, an illustration of the estimated marginal means is presented in Figure 2.3. Participants with average or below-average prior efficacy beliefs in the unblinded condition provided lower study quality evaluations than participants with equivalent beliefs in the blinded condition. This bias was consistent across levels of CRT scores for participants with average prior efficacy beliefs, yet it was nonsignificantly larger for those below-average prior efficacy beliefs who also had above-average CRT scores. In other words, participants for whom the presented results were the most politically-unfriendly exhibited slightly, but nonsignificantly, stronger biases in their evaluations if they were above-average in analytic thinking. Additionally, although participants with above-average prior efficacy beliefs (for whom the presented results were more politically-friendly) and below-average or average CRT scores did not significantly

differ in their quality evaluations across conditions, those with above-average prior efficacy beliefs and above-average CRT scores did significantly differ across conditions. Those more analytic participants evaluated the study as being of significantly higher quality when they made unblinded evaluations. In other words, highly analytic participants for whom the presented results were consistent with their prior efficacy beliefs provided positively biased judgments of study quality, but participants with similar prior efficacy beliefs who were lower in analytic thinking did not produce such positively biased judgments. In sum, being a more reflective thinker was associated with a slightly stronger biases in quality evaluations for unblinded participants, and this effect was stronger in participants for whom the presented results affirmed their prior efficacy beliefs.

Figure 2.3. Average Study Quality Evaluations by Condition and Prior Efficacy Beliefs and CRT scores

Error bars represent 95% CIs



Credibility impressions

Like in Studies 1a and 1b, the credibility impressions items had high internal reliability (Cronbach's $\alpha = 0.94$). I aggregated these items into a single measure, as preregistered. Participants had modestly positive credibility impressions on average ($M = 4.76$, $SD = 1.08$). Furthermore, participants in the blinded condition ($M = 4.89$, $SD = 1.04$) had higher average credibility impressions of the study than participants in the unblinded condition ($M = 4.63$, $SD = 1.11$), Welch's $t(907.87) = 3.62$, $p < .001$, Cohen's $d = 0.24$.

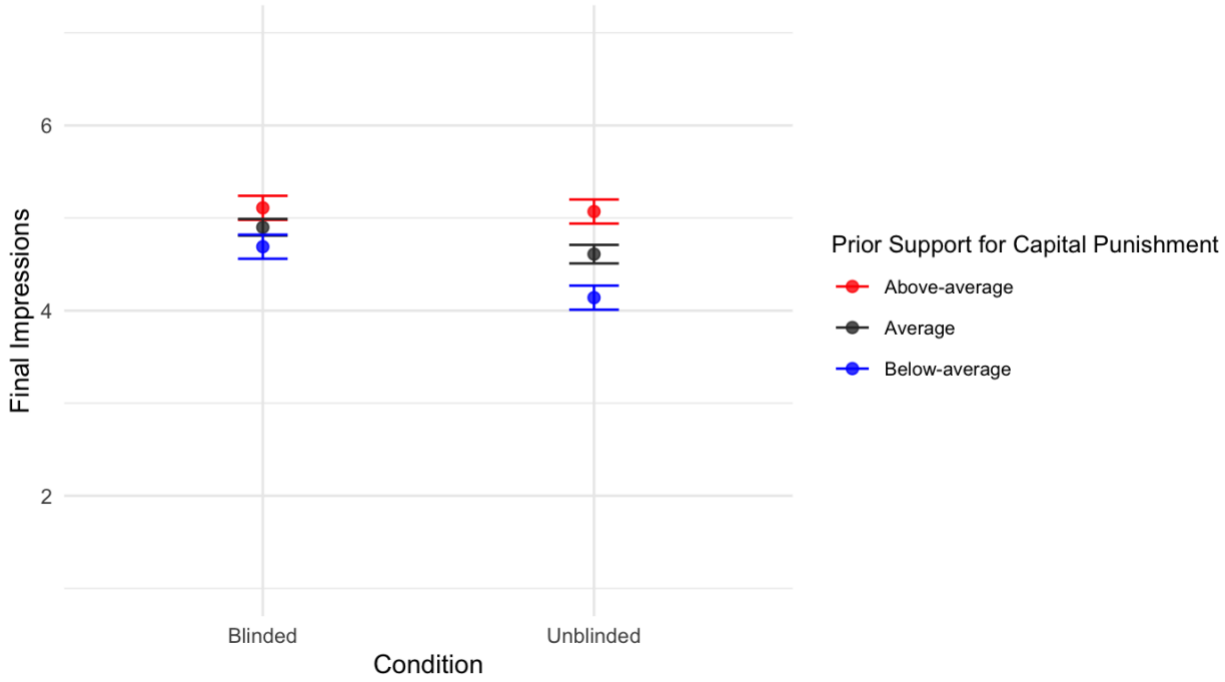
To further address Research Question 3, I conducted a series of linear regression and moderated mediation analyses to assess whether the effect of blinding on credibility impressions may have varied for participants with different prior beliefs, and to explore whether condition differences in study quality evaluations could help explain any such effects.

Influence of prior support

Starting with the linear regression model predicting credibility impressions from condition (dummy-coded, 0 = blinded), prior support (mean-centered), and their interaction, there was a significant main effect of condition, $F(1, 908) = 19.01$, $p < .001$, $\eta_p^2 = .02$, as well as a main effect of prior support $F(1, 908) = 19.05$, $p < .001$, $\eta_p^2 = .09$. As in Study 1a, these main effects were qualified by a significant interaction, $F(1, 908) = 14.82$, $p < .001$, $\eta_p^2 = .02$. Greater prior support for capital punishment was predictive of more positive credibility impressions across conditions, but this association was significantly stronger in the unblinded condition, $b = 0.23$, $SE = 0.02$, $p < .001$, 95% CI [0.19, 0.28], $\beta = 0.43$, than in the blinded condition, $b = 0.10$, $SE = 0.02$, $p < .001$, 95% CI [0.06, 0.15], $\beta = 0.19$. Furthermore, analyses of the estimated marginal means, illustrated in Figure 2.4, showed that participants with average and below-average prior support—the participants for whom the presented results were more politically-

unfriendly—had significantly more positive credibility impressions in the blinded condition than in the unblinded condition. Participants with above-average prior support for capital punishment, on the other hand, did not differ in their credibility impressions across conditions.

Figure 2.4. Average Final Impressions by Condition and Prior Support in Study 2
Error bars represent 95% CIs



The moderated mediation analysis using prior support as the moderator of the condition effects on credibility impressions, shown in Table 2.2, provided results consistent with those of the parallel analysis conducted in Study 1a. That is, while condition did not have significant indirect or direct effects on credibility impressions for participants with above-average prior support, it had countervailing indirect and direct effects on the credibility impressions of participants with average and below-average prior support. Those with average and below-average prior support provided lower study quality evaluations in the unblinded condition, which led them to form more negative credibility impressions of the study (e.g., find it less believable) in the unblinded condition as well. Yet there were also significant direct effects of for these participants, such that being in the unblinded condition caused a small increase in credibility

impressions. Nevertheless, the indirect effects were over three times as large as the direct effects, resulting in significant total effects of condition on credibility impressions. In total, these analyses showed that, as in Study 1a, participants who viewed politically-unfriendly information in the unblinded condition relied on their biased study quality evaluations when forming their credibility impressions, which caused them to have more negative credibility impressions than their blinded counterparts with similar prior support. Put differently, blinding participants quality judgments made them evaluate politically-unfriendly results more positively and, consequently, form more positive credibility impressions of the study.

Influence of prior efficacy beliefs

The linear regression and moderated mediation analyses using prior efficacy beliefs yielded substantively identical results as the analyses using prior support. Participants with average and below-average prior efficacy beliefs formed significantly more negative credibility impressions of the study in the unblinded condition than in the blinded condition, and this effect was explained by the influence of the negatively biased study quality evaluations those unblinded participants provided. Participants with above-average prior efficacy beliefs, on the other hand, did not significantly differ in their credibility impressions across conditions. The details of these models are provided in Appendix B.

Table 2.2. *Moderated mediation estimates of condition predicting credibility impressions by prior support for Study 2.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.80	0.08	[-0.96, -0.64]	-0.38	< .001
		Condition ⇒ Study Quality	-1.09	0.10	[-1.29, -0.90]	-0.51	< .001
		Study Quality ⇒ Credibility impressions	0.73	0.03	[0.67, 0.80]	0.74	< .001
	Direct	Condition ⇒ Credibility impressions	0.25	0.07	[0.11, 0.39]	0.12	< .001
	Total	Condition ⇒ Credibility impressions	-0.55	0.10	[-0.74, -0.37]	-0.26	< .001
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.38	0.05	[-0.48, -0.27]	-0.18	< .001
		Condition ⇒ Study Quality	-0.51	0.07	[-0.64, -0.37]	-0.24	< .001
		Study Quality ⇒ Credibility impressions	0.75	0.03	[0.70, 0.80]	0.75	< .001
	Direct	Condition ⇒ Credibility impressions	0.08	0.05	[-0.02, 0.17]	0.04	.11
	Total	Condition ⇒ Credibility impressions	-0.29	0.07	[-0.43, -0.16]	-0.14	< .001
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.06	0.07	[-0.08, 0.20]	0.03	.39
		Condition ⇒ Study Quality	0.08	0.09	[-0.10, 0.26]	0.04	.39
		Study Quality ⇒ Credibility impressions	0.77	0.04	[0.70, 0.84]	0.76	< .001
	Direct	Condition ⇒ Credibility impressions	-0.09	0.07	[-0.23, 0.04]	-0.04	.17
	Total	Condition ⇒ Credibility impressions	-0.03	0.10	[-0.22, 0.15]	-0.02	.72

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Updating support and efficacy beliefs

Finally, to address Research Question 3, I conducted a series of analyses to assess whether participants reported changing their support for capital punishment or beliefs about the efficacy of capital punishment. As in Studies 1a and 1b, I focused on the measures actual belief change (Time 2 beliefs – Time 1 beliefs) for both support and efficacy beliefs (analyses of the reported change in beliefs items are reported in Appendix B).

Updating support beliefs

In the linear regression model predicting change in support for capital punishment, the intercept (b = 0.14, SE = 0.04, 95% CI [0.05, 0.23]) showed that participants had a very small increase in their support in the blinded condition. There was not a significant main effect of condition, $F(1, 908) = 0.93$, $p = .34$, $\eta_p^2 = .00$, but there was a significant main effect of prior support, $F(1, 908) = 25.34$, $p < .001$, $\eta_p^2 = .03$, such that participants who previously supported capital punishment changed their support beliefs less on average, b = -0.11, SE = 0.02, 95% CI [-0.15, -0.07], β = -0.23. There was not a significant interaction between condition and prior support, $F(1, 908) = 1.94$, $p = .16$, $\eta_p^2 = .00$, so the lack of influence of blinding on changes in support beliefs was consistent across levels of prior support.

However, a moderated mediation analysis showed that the null effect of condition on change in support resulted from different combinations of influences for participants with different prior beliefs. These estimates are provided in Table 2.3. There were no significant indirect, direct, or total effects for participants with above-average prior support (participants for whom the results were politically-friendly), but there were significant indirect and direct effects for participants with average and below-average prior support. For those with average and below-average prior support, the indirect effect indicated that being in the unblinded condition

caused participants to lower their study quality evaluations relative to their blinded counterparts, which resulted in decreased change in support for those unblinded participants. Yet being in the unblinded condition also directly caused a small increase in change in support for these participants, canceling out the influence of the negative indirect effect for both groups. Thus, while the total impact of the manipulation on belief updating did not differ across levels of prior support, the ways in which the manipulation influenced belief updating differed as a function of the alignment of the results with participants' prior beliefs.

Updating efficacy beliefs

The intercept in the model predicting change in efficacy beliefs (b = 0.68, SE = 0.06, 95% CI [0.56, 0.79]) showed that participants in the blinded condition generally came to believe that capital punishment is more effective than they previously thought. There was not a significant main effect of condition, $F(1, 908) = 1.31$, $p = .25$, $\eta_p^2 = .00$, but there was an effect of prior efficacy beliefs, $F(1, 908) = 150.28$, $p < .001$, $\eta_p^2 = .14$, such that participants who previously believed that capital punishment is effective generally changed their efficacy beliefs less, b = -0.36, SE = 0.03, 95% CI [-0.42, -0.30], β = -0.50. The interaction between prior efficacy beliefs and condition was not significant, $F(1, 908) = 3.79$, $p = .052$, $\eta_p^2 = .00$, indicating that the null effect of condition on change in efficacy beliefs was consistent across levels of prior efficacy beliefs.

Nevertheless, the moderated mediation analysis, presented in Table 2.4, exposed differences in the effects of condition on change in efficacy beliefs that depended on participants' prior efficacy beliefs. As in the analysis of change in support, there were indirect effects of condition on change in efficacy beliefs for participants with average and below-average prior efficacy belief (those for whom the presented results were more politically-

unfriendly), such that being in the blinded condition increased the amount they changed their efficacy beliefs via their higher quality evaluations. This was not the case for participants with above-average prior efficacy beliefs. Yet unlike in the analysis of change in support beliefs, the total effect of condition on change in efficacy beliefs was significant for participants with below-average prior efficacy beliefs. That is, accounting for the indirect effect that the blinding manipulation had on increasing these participants' quality judgments revealed that those for whom the results were most politically-unfriendly changed their beliefs significantly more in the blinded condition relative to the unblinded condition. There were no significant indirect, direct, or total effects for participants with above-average prior efficacy beliefs. Ultimately, blinding the quality evaluations of participants with below-average prior efficacy beliefs—whose prior beliefs were most in conflict with the presented evidence—made them update their beliefs in the direction of the presented evidence significantly more than their unblinded counterparts.

Table 2.3. *Moderated mediation estimates of condition predicting change in support beliefs by prior support beliefs for Study 2.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	-0.26	0.07	[-0.39, -0.13]	-0.14	< .001
		Condition ⇒ Study Quality	-1.09	0.10	[-1.29, -0.91]	-0.51	< .001
		Study Quality ⇒ Change in Support	0.24	0.05	[0.13, 0.35]	0.27	< .001
	Direct	Condition ⇒ Change in Support	0.23	0.11	[0.02, 0.43]	0.12	.029
	Total	Condition ⇒ Change in Support	-0.03	0.09	[-0.20, 0.15]	-0.01	.76
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	-0.09	0.02	[-0.13, -0.05]	-0.05	< .001
		Condition ⇒ Study Quality	-0.51	0.06	[-0.63, -0.38]	-0.24	< .001
		Study Quality ⇒ Change in Support	0.18	0.03	[0.11, 0.25]	0.20	< .001
	Direct	Condition ⇒ Change in Support	0.19	0.07	[0.05, 0.32]	0.10	.007
	Total	Condition ⇒ Change in Support	0.06	0.06	[-0.06, 0.18]	0.03	.33
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Support	0.01	0.01	[-0.01, 0.03]	0.00	.39
		Condition ⇒ Study Quality	0.08	0.09	[-0.09, 0.25]	0.04	.38
		Study Quality ⇒ Change in Support	0.11	0.03	[0.05, 0.18]	0.13	< .001
	Direct	Condition ⇒ Change in Support	0.14	0.08	[-0.01, 0.28]	0.07	.066
	Total	Condition ⇒ Change in Support	0.15	0.09	[-0.03, 0.32]	0.08	.095

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 2.4. *Moderated mediation estimates of condition predicting change in efficacy beliefs by prior efficacy beliefs for Study 2.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	*β*	*p*
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	-0.47	0.08	[-0.62, -0.32]	-0.17	< .001
		Condition ⇒ Study Quality	-1.11	0.10	[-1.30, -0.92]	-0.52	< .001
		Study Quality ⇒ Change in Efficacy	0.42	0.06	[0.31, 0.53]	0.33	< .001
	Direct	Condition ⇒ Change in Efficacy	0.21	0.13	[-0.04, 0.46]	0.08	.10
	Total	Condition ⇒ Change in Efficacy	-0.25	0.11	[-0.47, -0.03]	-0.09	.029
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	-0.17	0.03	[-0.23, -0.11]	-0.06	< .001
		Condition ⇒ Study Quality	-0.49	0.06	[-0.61, -0.37]	-0.23	< .001
		Study Quality ⇒ Change in Efficacy	0.35	0.04	[0.27, 0.42]	0.27	< .001
	Direct	Condition ⇒ Change in Efficacy	0.13	0.09	[-0.04, 0.29]	0.05	.14
	Total	Condition ⇒ Change in Efficacy	-0.09	0.08	[-0.25, 0.06]	-0.03	.25
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Change in Efficacy	0.03	0.02	[-0.01, 0.08]	0.01	.15
		Condition ⇒ Study Quality	0.12	0.09	[-0.05, 0.29]	0.06	.15
		Study Quality ⇒ Change in Efficacy	0.27	0.05	[0.17, 0.36]	0.21	< .001
	Direct	Condition ⇒ Change in Efficacy	0.04	0.11	[-0.18, 0.25]	0.01	.72
	Total	Condition ⇒ Change in Efficacy	0.06	0.11	[-0.16, 0.29]	0.02	.57

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Discussion

The results of Study 2 were almost entirely in line with those of Study 1a. Blinding participants' quality evaluations provided accuracy-motivated evaluative baselines across levels of prior belief against which to compare unblinded participants' evaluations. While blinded participants' prior support and prior efficacy beliefs did not significantly influence their quality evaluations, these beliefs did influence the evaluations of unblinded participants. This showed that these unblinded participants provided biased evaluations of study quality, as the congeniality of the results with one's prior beliefs would not impact how a purely accuracy-motivated evaluated the quality of new information—as was the case in the blinded condition. These biases were most easily detected in participants who came into the study more opposed to capital punishment, for whom the presented results were the most politically-unfriendly. Supporting my hypothesis for Research Question 1, partisans' prior beliefs biased their quality evaluations in Study 2.

As in Study 1a, and supporting my hypothesis regarding Research Question 2, both prior support and prior efficacy beliefs were found to bias unblinded participants' quality judgments. Nonetheless, unlike in Study 1a, both prior support and prior efficacy beliefs significantly predicted quality evaluations when accounting for shared variance between the two belief measures. Indeed, including both belief measures in the same model only modestly increased the amount of variance in quality evaluations that could be explained, suggesting that the measures were largely capturing similar variance in participants' biased evaluations. This suggested that the directional motives underlying these biased evaluations may manifest their influence via various forms of belief. In other words, while researchers have long argued that people's feelings (e.g., support beliefs) can act as the conduit through which directional motivations bias thinking

(Ditto et al., 2009; Ditto et al., 2019a), these results indicate that people's expectations (e.g., efficacy beliefs) can be a similar vehicle through which directional motivations influence evaluations of new information. The biasing influence that partisans' expectations can have on their evaluations has not been widely recognized in past research (for an exception, see Druckman & McGrath, 2019).

Despite the robust evidence of partisan bias in this study, there was little indication that the biased evaluations were affectively-driven. Accounting for participants positive and negative affect explained very little variance in their quality judgments. However, participants may not have been willing to share their true emotional reactions to the presented information. If this is true, then affective reactions to the results may account for more of the biasing effects of participants' prior beliefs than was captured in this study, as has been documented in prior research (Munro et al., 2002). Indeed, there were some indications that moral conviction, which is indicative of emotional investment (Skitka & Wisneski, 2011), exacerbated the observed biases, yet these effects were small and generally nonsignificant. Thus, given the absence of robust affective mediation or moderation, the present results suggest that the observed biases were more cognitively-driven. Further supporting this contention, participants higher in analytic thinking demonstrated slightly stronger biases than less analytic individuals with equivalent efficacy beliefs. Nevertheless, analytic thinking did not similarly exacerbate the biasing influence of support beliefs on quality evaluations, and the moderating effect of analytic thinking on the biases stemming from prior efficacy beliefs was small. In sum, although the psychological mechanisms underlying the observed biases remain unclear, it appeared that these biases were not reducible to affective mechanisms.

Lastly, in relation to Research Question 3, unblinded participants' biased evaluations accounted for condition differences in their downstream credibility impressions and belief updating. Participants who knew the study results were politically-unfriendly when they evaluated its methods formed more negative credibility impressions and updated their efficacy beliefs less than their blinded counterparts. Although similar patterns emerged for changes in support beliefs, the effect was much smaller and not significant. Overall, Study 2 demonstrated that partisans who made unblinded judgments of scientific information provided directionally biased evaluations, and these biases caused participants for whom the presented results were politically-unfriendly to be less open to updating their beliefs after considering the presented evidence.

STUDY 3

While Study 2 addressed some of the limitations of Studies 1a and 1b, there were additional shortcomings of the previous studies that I aimed to address in Study 3. First, the stimuli used in the previous studies were generally politically-unfriendly to more liberal participants. Although the totality of current evidence suggests that partisan bias is equivalent for liberals and conservatives (Ditto et al., 2019a), some researchers suggest that conservatives are more prone to partisan biases than liberals (Adorno et al., 1950; Baron & Jost, 2019). To test for asymmetries in partisan bias more directly, which relates to Research Question 2 about the drivers of partisan bias, I developed two sets of materials for Study 3. One set of materials showed liberal-friendly results, and the other showed conservative-friendly results. Comparing partisans' evaluations across these materials afforded more direct tests of the symmetry and asymmetry hypotheses.

Second, while Studies 1a and 2 yielded evidence that clearly aligned with my hypotheses for Research Questions 1 and 2, the results of Study were less clear. This was likely due, at least in part, to having too small of a sample in Study 1b, but it is also possible that partisan bias is only detectable in evaluations of certain topics, like capital punishment. To test whether evaluations of scientific information about other topics may yield similar biases as Studies 1a and 2, I developed stimuli regarding a topic of political interest that does not have direct policy implications—partisan bias. That is, the materials for Study 3 were based on our meta-analysis of partisan bias (Ditto et al., 2019a), and participants were informed that the results showed that either liberals or conservatives are more biased, depending on which materials they were presented. Using this topic also allowed me to test whether feelings and beliefs other than policy support and efficacy beliefs—namely, feeling about partisan ingroups and outgroups and beliefs

about partisan bias—may bias evaluations of scientific quality. The items capturing prior feelings about partisan ingroups and outgroups (i.e., partisan feelings) were conceptualized as indexing more affectively-driven processes, and the measure of prior beliefs about partisan bias (i.e., prior bias beliefs), were conceptualized as indexing more cognitively-driven processes.

Third, in relation to Research Question 2, I included additional individual difference measures to test for moderation of the confirmatory analyses. I measured participants confidence in their prior bias beliefs to test whether more confident participants displayed stronger evaluative biases. I also included the measures of analytic thinking used in Study 2 and a measure of intellectual humility to further investigate whether cognitive and metacognitive dispositions may exacerbate (or mitigate) the biasing influence of partisans' prior beliefs on their quality evaluations. Additionally, I included an exploratory measure of social concern (i.e., how liberal or conservative one's social environment is) to see whether merely inhabiting a more polarized environment could bias participants' quality evaluations, which would accord with some motivated accounts of partisan reasoning (Kahan, 2016). Lastly, I measured participants' strength of partisan identification to explore whether partisans' with stronger partisan considerations may demonstrate stronger biases than their less-identified peers.

Nevertheless, the primary hypotheses of Study 3 were similar to those in the previous studies. I hypothesized that participants' prior beliefs would significantly predict quality evaluations only in the unblinded conditions (addressing Research Question 1) and that both partisan feelings and prior bias beliefs would significantly bias unblinded participants' quality evaluations (addressing Research Question 2). Additionally, I preregistered analyses to examine how the blinding manipulation and participants' prior beliefs influenced their credibility impressions and belief updating (addressing Research Question 3).

Method

Participants

A power analysis indicated that I would need to recruit at least 1700 participants to have 0.95 power to detect small (f = 0.12) three-way interactions. To collect a sample with sufficient political diversity, I used Prolific's U.S. Political Affiliation demographic screeners to recruit a stratified random sample. Expecting to exclude several participants, I aimed to recruit 2000 participants ($n_{Democrats}$ = 700, $n_{Republicans}$ = 700, $n_{Independents}$ = 600) through Prolific.

Ultimately, 2002 participants were recruited in January of 2022. The participants ranged in age from 18-84 years (M = 36.83, SD = 14.77), ranged in yearly household income from less than $5,000 to over $175,000 (*Mdn* = $60,000 - $74,999) and were mainly White (77%), female (60%), and college-educated (37%). Based on a 7-pt. measure of party affiliation I included in the demographics section (1 = *Strong Republican*, 2 = *Republican*, 3 = *Lean Republican*, 4 = *Neither Republican or Democrat*, 5 = *Lean Democrat*, 6 = *Democrat*, 7 = *Strong Democrat*), I recruited 851 Democrats, 753 Republicans, and 398 Independents[5].

Procedure and Measures

After consenting, participants were asked, "In general, how negatively or positively do you feel towards the following groups of people?" and responded on a sliding scale (-10 = *Extremely negative*, 0 = *Neutral*, 10 = *Extremely positive*) about Democrats, Republicans, liberals, and conservatives in a randomized order. These items served as the measure of partisan feelings. Participants were then asked, "To what extent do you believe that liberals are more politically biased than conservatives (or that conservatives are more politically biased than

[5] Based on Prolific's demographic screeners, I recruited 702 Democrats, 699 Republicans, and 601 Independents/Others/None. It is unclear why there was a discrepancy between participants' screener responses and their responses to my measure of party affiliation, but more participants identified with a political party in my survey than on Prolific's screening question.

liberals)?" and responded on a 7-pt. scale (1 = *Liberals are much more biased than Conservatives*, 4 = *Liberals and Conservatives are equally biased*, 7 = *Conservatives are much more biased than Liberals*). This item served as the measure of prior beliefs. Participants also indicated their confidence in their prior beliefs by responding to the following question on a 7-pt. scale: "How confident are you in the belief that you reported (about political bias) in the previous question?" (1 = *Not confident at all*, 4 = *Moderately confident*, 7 = *Extremely confident*).

On the next page, participants completed four additional items as an exploratory measure of social concern. The prompt to these items read as follows:

> We are going to show you a series of statements a person could advocate for. For each statement, say how negative you think your community (the people you interact with, friends, family, etc.) would react toward you if you advocated for such a statement (for example, on social media, or during Thanksgiving dinner, etc.). In other words, if you were to make such a statement in public, how negatively (or positively) would people react toward you?

Participants then responded on sliding scales (-10 = *Extremely negative*, 0 = *Neutral*, 10 = *Extremely positive*) to the following statements: "Liberals are more biased than conservatives," "Conservatives are more biased than liberals," "Liberals are the ones responsible for the dysfunction in our current politics," and "Conservatives are the ones responsible for the dysfunction in our current politics." The responses to the conservative-focused items were reversed-scored, and the four items were aggregated into a composite social concern measure.

Participants were then randomly assigned to one of four experimental conditions in a 2 (Condition: blinded or unblinded) by 2 (Materials: liberal-friendly or conservative-friendly) between-subjects design. As in my previous experiments, all participants read a brief introduction and methods description of the focal study, which was a meta-analysis on partisan bias:

> A controversial public question in recent years has been whether liberals and conservatives differ in their tendencies to process information in a biased way. Some researchers suggest that conservatives demonstrate more bias than liberals, whereas others deny this and believe that partisans on both sides are equally biased or that liberals demonstrate more bias. A recent research effort attempted to shed light on this controversy.
>
> Researchers from the University of California, Irvine (Ditto et al., 2019) published a study in *Perspectives on Psychological Science* that looked at differences in political bias. The researchers conducted a meta-analysis of 51 experimental studies which involved over 18,000 total participants. The authors only looked at studies conducted in the United States that examined a specific form of political bias - the tendency to evaluate otherwise identical information more favorably when it supports one's political beliefs than when it challenges those beliefs. For example, one study that was analyzed looked at whether liberals and conservatives supported the same welfare policy more when it was endorsed by a particular political group (Democrats or Republicans). The researchers tested for differences in political bias by examining whether the average bias score for liberals significantly differed from the average bias score for conservatives across the 51 studies.

Participants in the unblinded conditions were only required to stay on the page containing this description for 30 seconds before they could proceed to the net page. Alternatively, those in the blinded conditions were required to stay on the page for 30 seconds but were also presented the study quality items immediately after the methods description. The wording of the study quality items for this study are shown in Table 3.1.

On the next page, all participants were presented with the full write-up of the study, which included the introduction and methods descriptions and a brief description of the results. All participants were required to stay on this page for at least 30 seconds before proceeding. Those presented with conservative-friendly materials read the following: "Their results, as illustrated in the figure below, showed that liberals were approximately twice as biased as conservatives on average. The researchers concluded that liberals demonstrated more political bias than conservatives in the studies they examined." Participants presented with liberal-friendly materials read an otherwise identical results description indicating that conservatives

demonstrated more political bias than liberals (the figures accompanying these descriptions are presented in Appendix C). Participants in the unblinded conditions then completed the study quality measures, while participants in the blinded conditions were not presented any items before proceeding.

All participants were then presented with the manipulation check and the credibility impressions items from the previous studies. For the attention check, participants responded to the question "What did the results of the study you read about show?" on a 7-pt. scale (1 = *Liberals are much more biased than Conservatives*, 4 = *Liberals and Conservatives are equally biased*, 7 = *Conservatives are much more biased than Liberals*). The credibility impression items were identical to those in previous studies other than altering details to match the presented stimuli (e.g., the university where the research was conducted). On the following page, participants were asked to complete the set of partisan feeling thermometers, the item about beliefs about partisan bias, and the item about their confidence in their beliefs about partisan bias for a second time. These items were nearly identical to those presented at the beginning of the study and served as the second time point for actual belief change measures.

Next, participants completed the Cognitive Reflection Task 1 and 2 (CRT; Frederick, 2005; Thomson & Oppenheimer, 2016) and the General Intellectual Humility Scale (GIHS; Leary et al., 2017) in randomized order. Participants then completed demographic information about their age, sex, ethnicity, income, education, political orientation (social and economic, separately), and political party affiliation. Unlike in my previous studies, participants who indicated they at least leaned toward a party were subsequently given a strength of partisan identification measure (Bankert et al., 2017). Finally, participants were given the opportunity to

share any thoughts or feelings about the study in an open-response question before being presented with the debriefing.

Table 3.1. *Study Quality Items for Study 3.*

1. How well do you think the researchers measured participants' levels of political bias? (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)
2. How well does assessing judgments of otherwise identical information when it supports (vs. challenges) one's political beliefs measure participants' actual levels of political bias? (1 = *Very poorly,* 4 = *Neither well nor poorly,* 7 = *Very well*)
3. How appropriate was the sample size (51 studies, around 18,000 participants) for the study the researchers conducted? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
4. How appropriate was it to analyze these data (Americans' responses to political information) to address the research question? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
5. How appropriate was the approach used by the researchers in this study (a meta-analysis of experimental studies) for answering the research question? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
6. How appropriate of comparison groups (liberals and conservatives) did the researchers use in this study? (1 = *Very inappropriate,* 4 = *Neither appropriate nor inappropriate,* 7 = *Very appropriate*)
7. Are the data from this study helpful for answering the research question? (1 = *Definitely unhelpful,* 4 = *Neither helpful nor unhelpful,* 7 = *Definitely helpful*)
8. On the scale below, please indicate how valid you would find the results of the above study. (1 = *Very invalid,* 4 = *Neither valid nor invalid,* 7 = *Very valid*)

Results

Following my preregistration, I excluded participants who failed our manipulation check[6] or finished the survey in less than four minutes. This resulted in a sample of 1762 participants for the confirmatory analyses (including the full sample did not substantively alter the results). The distribution of participants across the four experimental conditions was roughly equivalent ($n_{blinded/lib\text{-}friendly}$ = 468, $n_{blinded/con\text{-}friendly}$ = 428, $n_{unblinded/lib\text{-}friendly}$ = 455, $n_{unblinded/con\text{-}friendly}$ = 411). The final sample consisted of 762 Democrats, 649 Republicans, and 351 Independents.

On average, participants felt slightly warm toward liberals and Democrats (M = 0.56, SD = 6.15) and slightly cold toward conservatives and Republicans (M = -1.50, SD = 6.23). Following my preregistration, I first conducted analyses using these composites separately, but I used a four-item composite feeling thermometer measure for the analyses presented below (the results did not substantively differ when using either of the original two-item composites). The four-item measure aggregated participants' feelings toward liberals and Democrats with reverse-scored items of their feelings toward conservatives and Republicans (Cronbach's α = 0.91). For this combined feeling thermometer composite, positive scores indicated feeling more positively about liberals/Democrats than conservatives/Republicans, and negative scores indicated feeling more positively about conservatives/Republicans than liberals/Democrats. Participants reported feeling roughly equal about liberals and conservatives on average (M = 0.84, SD = 5.28)[7]. Additionally, the exploratory measure of social concern displayed suitable reliability (Cronbach's α = 0.87). For the social concern composite measure, negative scores indicated a

[6] Participants who failed the manipulation check either reported that a) the presented results showed liberals and conservatives are equally biased, or b) that the results showed the opposite group was biased from what was in the stimuli.

[7] Democratic participants felt significant more positive about liberals/Democrats (M = 5.40, SD = 2.63), than did Republicans (M = -4.52, SD = 3.11),Welch's $t(1275.31) = 64.04$, $p < .001$, indicating that our self-identified partisan participants truly held different partisan allegiances.

more liberal-friendly social environment, and positive scores indicated a more conservative-friendly social environment. Participants reported inhabiting politically-neutral social environments on average (M = -0.52, SD = 4.66).

Participants also reported that liberals and conservatives were equally biased on average (M = 4.09, SD = 1.59), and they tended to be moderately confident in their prior bias beliefs (M = 5.19, SD = 1.40). As in the previous studies, indices of skewness and kurtosis indicated that the distribution of these responses were sufficiently normal for parametric analyses (Field et al., 2012).

Study quality evaluations

Once again, the study quality items had high internal reliability (Cronbach's $\alpha = 0.91$) and were thus averaged into a composite measure, as preregistered. On average, participants rated the study as being of decent quality (M = 5.28, SD = 1.01). To address Research Questions 1 and 2 about the existence and drivers of partisan bias, I conducted a series of linear regression analyses using the GAMLj package in jamovi.

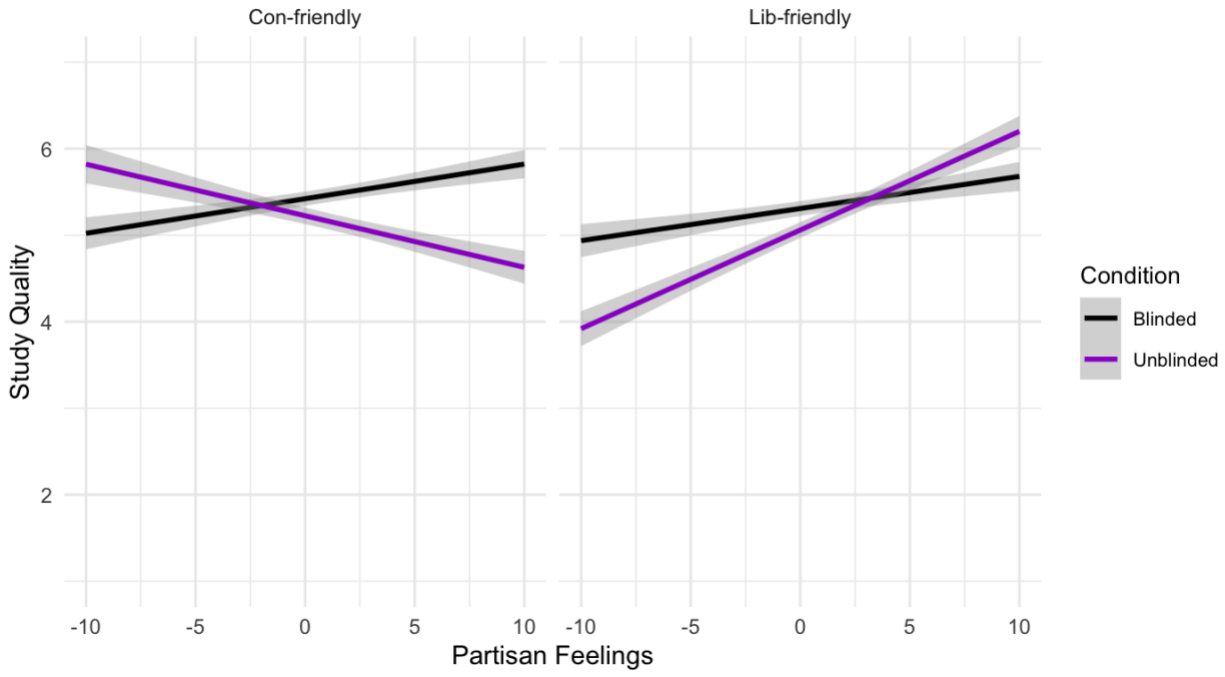
Influence of partisan feelings

First, I constructed a model predicting study quality evaluations from condition (dummy-coded, 0 = blinded), materials (dummy-coded, 0 = conservative-friendly), partisan feelings (mean-centered), and the two- and three-way interactions between these variables interaction. Omnibus tests showed that the overall model was significant, $F(7, 1754) = 44.74$, $p < .001$, $\eta_p^2 = .15$, as were the main effects of condition, $F(1, 1754) = 19.02$, $p < .001$, $\eta_p^2 = .01$, and partisan feelings, $F(1, 1754) = 21.83$, $p < .001$, $\eta_p^2 = .01$, but not of materials, $F(1, 1754) = 3.45$, $p = .064$, $\eta_p^2 = .00$. Simple effects analyses showed that, unexpectedly, partisan feelings were associated with study quality evaluations in the blinded conditions for both the conservative-

friendly materials, b = 0.04, SE = 0.01, p < .001, 95% CI [0.02, 0.06], β = 0.21, and the liberal-friendly materials, b = 0.04, SE = 0.01, p < .001, 95% CI [0.02, 0.05], β = 0.19. Blinded participants with more liberal-leaning feelings tended to have higher study quality evaluations, despite not knowing the study results when making these evaluations. This suggests that our measure of partisan feelings also captured aspects of participants' epistemic preferences or general trust in science, such that blinded liberal-leaning participants found the study's methodology to be sounder than blinded conservative-leaning participants. Nevertheless, to determine whether participants in the unblinded conditions were biased by their partisan feelings when making quality evaluations, the critical test was the three-way interaction, which tested whether the slopes between the blinded and unblinded conditions significantly differed within the same materials. This three-way interaction term was indeed significant, F(1, 1754) = 109.24, p < .001, η_p^2 = .06, illustrated in Figure 3.1. The influence that partisan feelings had on quality evaluations was significantly larger in the unblinded conditions in both the conservative-friendly, b = -0.06, SE = 0.01, p < .001, 95% CI [-0.08, -0.04], β = -0.31, and liberal-friendly materials, b = 0.11, SE = 0.01, p < .001, 95% CI [0.10, 0.13], β = 0.60. Unblinded participants with more liberal-leaning feelings evaluated the study much more positively than those with more conservative-leaning feelings when reading liberal-friendly results, yet unblinded participants with more liberal-leaning feelings evaluated the study much more *negatively* than those more conservative-leaning participants when reading conservative-friendly results. Thus, knowing the results of the study amplified the association between partisan feelings and study quality in the liberal-friendly condition, and it flipped the association between partisan feelings and study quality in the conservative-friendly condition. The difference between the unblinded and blinded

estimates within the same materials condition is the magnitude of directional bias that partisan feelings exerted on study quality evaluations.

Figure 3.1. Study Quality Evaluations by Partisan Feelings, Materials, and Condition in Study 3
Error bands represent standard errors



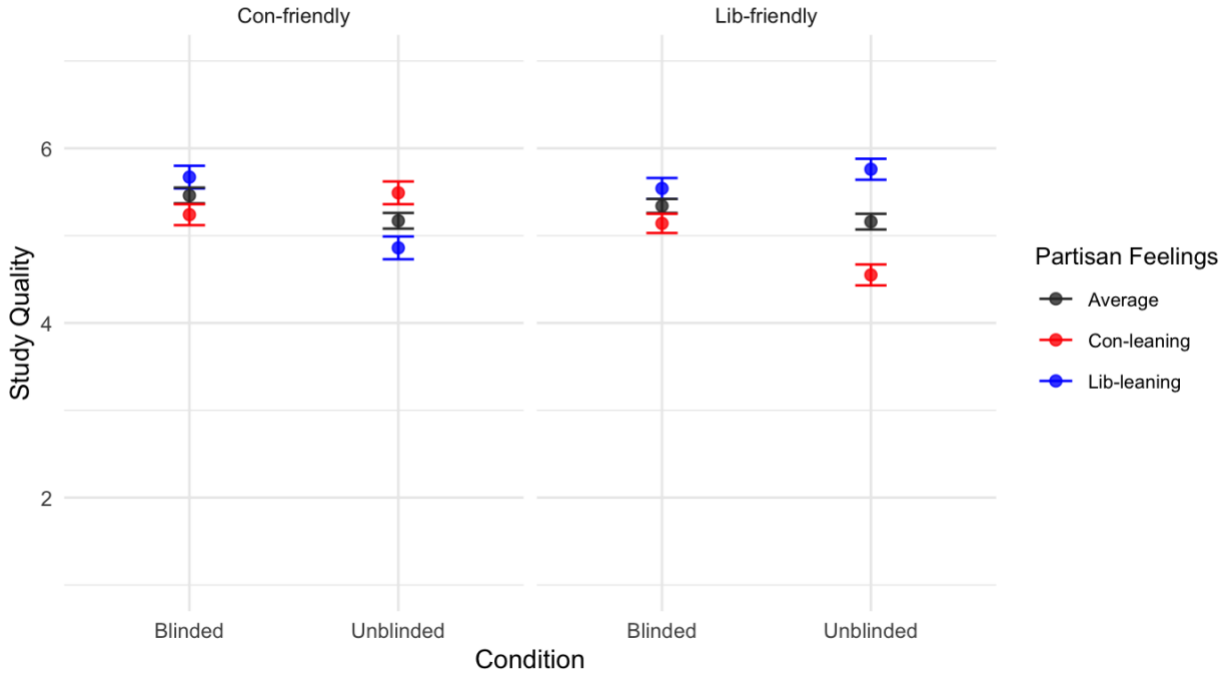
To further address Research Question 1, I compared the estimated marginal means of participants with conservative-leaning (below-average; $M - 1SD$), average, and liberal-leaning (above-average; $M + 1SD$) partisan feelings across conditions. Figure 3.2 depicts these estimated marginal means. Starting with the conservative-friendly materials, participants with conservative-leaning feelings provided slightly (but nonsignificantly) more positive study quality evaluations when they evaluated the study while knowing its results. Those with liberal-leaning feelings, on the other hand, provided significantly lower evaluations in the unblinded condition when they were presented with the conservative-friendly results. The opposite pattern emerged in participants who evaluated the liberal-friendly materials: those with conservative-leaning feelings provided significantly more positive study quality evaluations in the blinded condition, and liberal-leaning participants provided slightly (but nonsignificantly) more positive quality

evaluations in the unblinded condition. Participants with average partisan feelings were generally more consistent in their study quality evaluations across conditions, though they provided slightly more positive quality evaluations in the blinded conditions.

Notably, while the condition differences between liberal-leaning participants evaluations of the conservative-friendly materials (*b* = -0.81, 95 % CI[-0.99,-0.63]) was slightly larger than the condition differences between conservative-leaning participants of the liberal-friendly materials (*b* = -0.59, 95 % CI[-0.76,-0.42]), this difference was not significant. In relation to Research Question 2, this suggested that the magnitude of partisan bias in the evaluation of politically-unfriendly information was symmetrical for liberals and conservatives.

Figure 3.2. Average Study Quality Evaluations by Condition, Materials, and Partisan Feelings in Study 3

Error bars represent 95% CIs



In sum, these results further demonstrated that partisans who provided unblinded quality evaluations were biased by their partisan feelings when making those judgments. This directional bias emerged most clearly in partisans' evaluations of politically-unfriendly information. Compared to affectively-similar partisans who provided blinded evaluations of politically-

politically-unfriendly information, partisans who provided unblinded quality evaluations of politically-unfriendly information provided significantly lower quality ratings.

Influence of prior bias beliefs

Next, I constructed a linear regression models predicting study quality evaluations using condition (dummy-coded, 0 = blinded), materials (dummy-coded, 0 = conservative-friendly), prior bias beliefs (mean-centered), and the two- and three-way interactions between these variables interaction. The results of this analysis were nearly identical to those using partisan feelings and are depicted in Figures 3.3 and 3.4 (full statistical details of the model are provided in Appendix C). The results showed that, like partisan feelings, participants' prior bias beliefs directionally influenced their unblinded study quality evaluations, particularly for information that was inconsistent with their prior beliefs. Compared to partisans with similar prior bias beliefs who provided either unblinded evaluations of politically-friendly results or blinded evaluations of the study, partisans who provided unblinded quality evaluations of politically-unfriendly results evaluated the study more negatively.

Figure 3.3. Average Study Quality Evaluations by Prior Bias Beliefs, Materials, and Condition in Study 3
Error bands represent standard errors

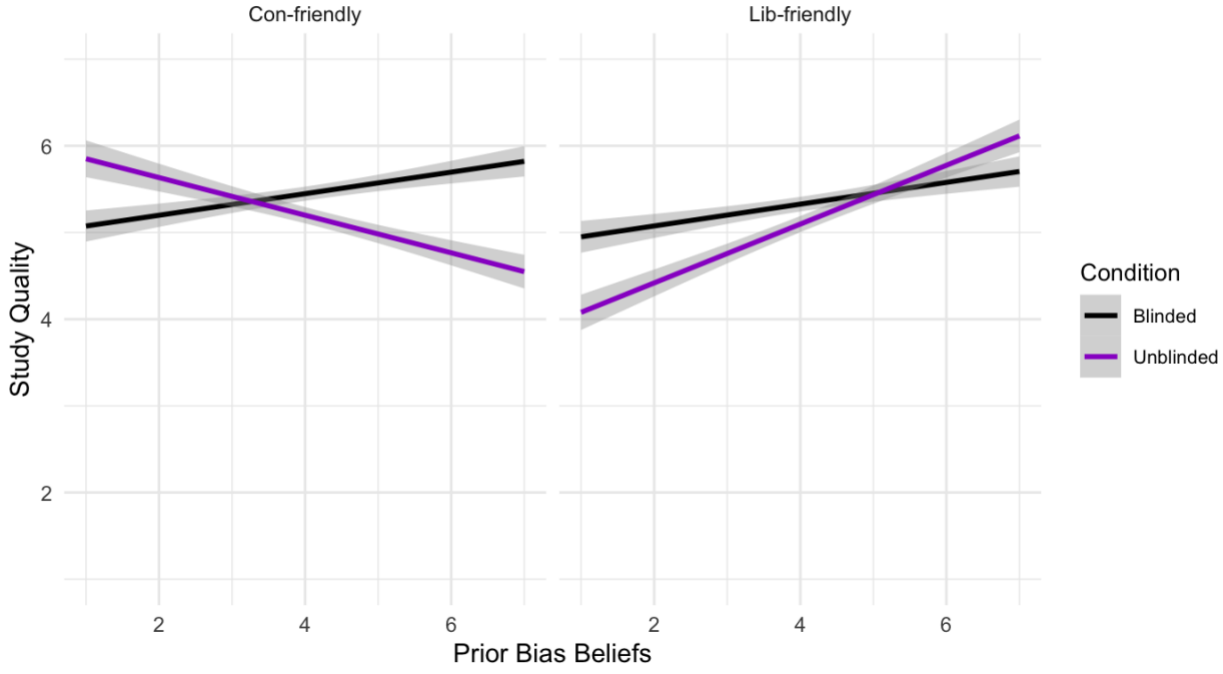
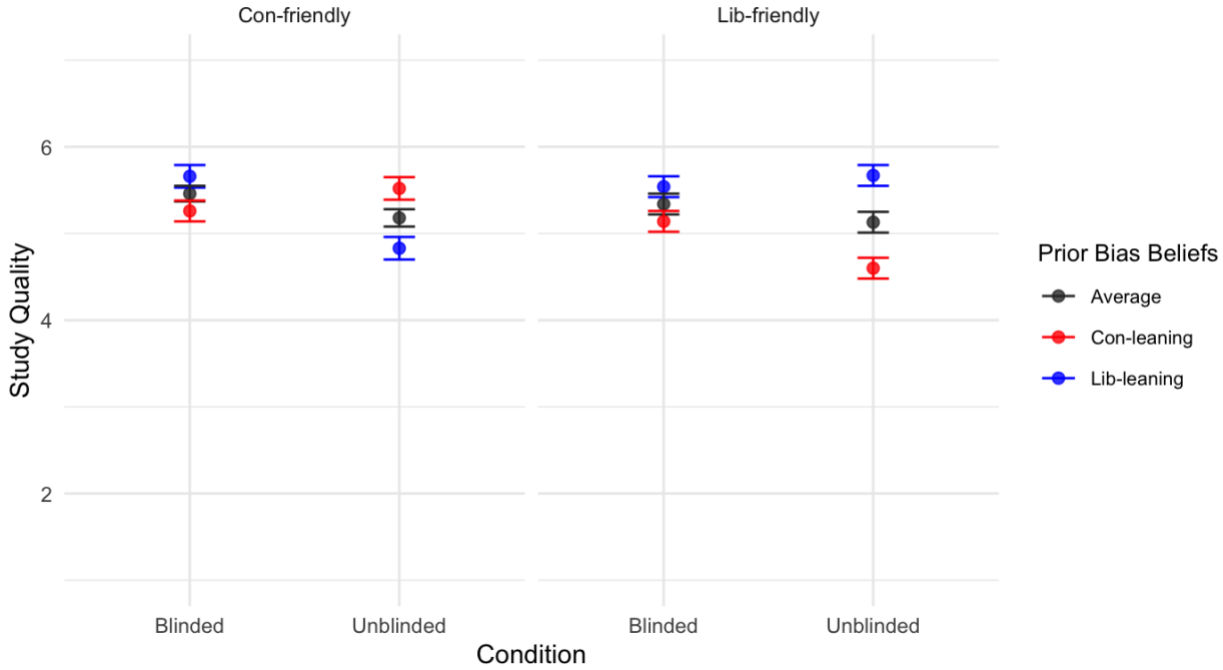


Figure 3.4. Average Study Quality Evaluations by Condition, Materials, and Prior Bias Beliefs in Study 3
Error bars represent 95% CIs



Relative influence of partisan feelings and prior bias beliefs

To further address Research Question 2 about the drivers of partisan bias, I constructed a regression model using condition, materials, both prior belief measures (partisan feelings and prior bias beliefs). The key terms in this combined model were the two three-way interaction terms, which indicated the influence that each prior belief measure had on quality evaluations when accounting for the shared variance between the two measures.

The results of this combined model indicated that the effects of partisan feelings and prior bias beliefs were both independent and intertwined. Omnibus tests indicated that the overall model was significant, $F(11, 1750) = 31.99$, $p < .001$, $\eta_p^2 = .17$. Notably, the total variance explained by this combined model ($\eta_p^2 = 0.17$) was only slightly larger than the variance explained by the models that only included a single interaction term with either partisan feelings ($\eta_p^2 = 0.15$) or prior beliefs ($\eta_p^2 = 0.14$). This suggested that the models with single interaction terms accounted for most of the same variance. Further aligning with this assessment, the three-way interaction including partisan feelings, $F(3, 1750) = 14.05$, $p < .001$, $\eta_p^2 = .02$, and the three-way interaction including prior bias beliefs, $F(3, 1750) = 9.94$, $p < .001$, $\eta_p^2 = .02$, were both significant in this combined model, yet these interaction terms accounted for less total variance ($\eta_p^2 = .04$) than the same interaction terms in the models with only an interaction term including partisan feelings ($\eta_p^2 = .06$) or prior bias beliefs ($\eta_p^2 = .05$). Given the limited variance explained by the main effects and two-way interactions in this combined model ($\eta_p^2 = .03$), these results indicated that there was a substantial amount of variance explained by the shared influence of the two three-way interaction terms ($\eta_p^2 = .10$). In other words, most of the biasing effect that the two prior belief measures had on participants' quality evaluations was due to their shared influence, rather than their independent influence. Thus, as in Study 2, distinct sets of prior

beliefs produced distinguishable—yet highly overlapping—influences on participants' study quality evaluations, which resulted in biased evaluations in the unblinded conditions.

Exploratory analyses of study quality evaluations

Following my preregistration, I conducted exploratory analyses using the measure of social concern as a predictor of quality evaluations, and I ran separate analyses to test whether participants' strength of partisan identification (for those who identified as either Democrat or Republican) predicted study quality evaluations as well. The full set of exploratory analyses are presented in Appendix C. In brief, models using the social concern measure yielded results consistent with those using the partisan feelings and prior beliefs measures, and participants who were more strongly identified with a political party demonstrated significantly stronger biases in quality evaluations than more weakly identified partisans. These results provided further evidence that the observed biases were not driven merely by participants' expectations, for their social and political considerations biased their quality evaluations.

Moderation of study quality evaluations by individual difference measures

Next, to examine additional questions related to Research Question 2, I constructed a series of four linear regression models to test whether any of the individual difference measures attenuated (or exacerbated) the previously observed biases in quality evaluations. In these analyses, I used a single factor variable that combined condition and materials. This combined variable had four levels (blinded/conservative-friendly, blinded/liberal-friendly, unblinded/conservative-friendly, unblinded/liberal-friendly) and was used to simplify the interpretations of the results (the results are substantively identical when separating condition and materials into two separate factors, as was done in the prior analyses). Accordingly, these models included the combined condition variable (dummy-coded, 0 = blinded/conservative-

friendly), either the partisan feelings or prior bias beliefs variable (mean-centered), either CRT or GIHS scores (mean-centered), and the corresponding two-way and three-way interaction terms.

Moderation of partisan feelings effects

Analytic thinking. Starting with the measure of analytic thinking, omnibus tests showed that CRT scores did not have a significant main effect on study quality evaluations, $F(1, 1746) = 0.42$, $p = .52$, $\eta_p^2 = .00$. Neither the interaction between CRT and condition, $F(3, 1746) = 0.60$, $p = .62$, $\eta_p^2 = .00$, CRT and partisan feelings, $F(1, 1746) = 0.73$, $p = .39$, $\eta_p^2 = .00$, nor the three-way interaction was significant, $F(3, 1746) = 0.47$, $p = .70$, $\eta_p^2 = .00$. There was a significant interaction between partisan feelings and condition, $F(1, 1746) = 68.68$, $p < .001$, $\eta_p^2 = .11$, as observed in the previous analyses. However, these results indicated that analytic thinking did not increase or decrease the influence that partisan feelings had on participants' quality judgments across conditions.

Intellectual humility. Omnibus tests showed that there was a significant main effect of GIHS scores on study quality evaluations, $F(1, 1746) = 10.35$, $p = .001$, $\eta_p^2 = .01$, such that humbler participants tended to rate the study as being of higher quality, $b = 0.19$, $SE = 0.06$, 95% CI [0.07, 0.30], $\beta = 0.15$. However, beyond the previously documented two-way interaction between partisan feelings and condition, there were not significant interactions between GIHS and condition, $F(3, 1746) = 0.32$, $p = .81$, $\eta_p^2 = .00$, GIHS and partisan feelings, $F(1, 1746) = 3.40$, $p = .066$, $\eta_p^2 = .00$, nor a three-way interaction, $F(3, 1746) = 0.55$, $p = .65$, $\eta_p^2 = .00$. Overall, this suggested that, while more intellectual humility participants judged the study more positively, intellectual humility did not attenuate or exacerbate the biasing influence of partisan feelings on unblinded participants' quality evaluations.

Moderation of prior bias beliefs effects

Analytic thinking. Omnibus tests showed that CRT scores did not have a significant main effect on study quality evaluations, $F(1, 1746) = 0.33$, $p = .56$, $\eta_p^2 = .00$, nor did it significantly interact with condition or prior bias beliefs in either of the two-way interaction terms or the three-way interaction (all ps $\geq .66$). As such, I found no evidence of analytic thinking increasing or minimizing the influence of prior bias beliefs on study quality evaluations.

Intellectual humility. Although GIHS scores had a significant main effect on study quality evaluations, $F(1, 1746) = 10.10$, $p = .002$, t $\eta_p^2 = .01$, there were not significant two- or three-way interactions in this model either (all ps $\geq .15$). Thus, those higher in self-reported intellectual humility tended to evaluate the study more positively, $b = 0.19$, $SE = 0.06$, 95% CI [0.07, 0.30], $\beta = 0.15$, but these participants were equally biased by their prior bias beliefs when making unblinded quality judgments.

Ultimately, these analyses did not find evidence of analytic thinking exacerbating biased evaluations, as was found in Study 2, and intellectual humility did not exacerbate or attenuate the observed biases either.

Moderation of study quality evaluations by confidence in prior bias beliefs

To assess how participants' confidence in their prior beliefs about bias influenced their quality evaluations, I constructed a regression model using the combined condition variable (dummy coded, 0 = blinded/conservative-friendly), prior bias beliefs (mean-centered), the measure of confidence in prior bias beliefs (mean-centered), the three two-way interactions between these variables, and the three-way interaction term. The key term in this model was the three-way interaction, which indicated whether the extent to which prior bias beliefs influenced quality evaluations across conditions varied as a function of one's confidence in their priors (the results are substantively identical when including condition and materials as separate factors).

Omnibus tests showed that the overall model was significant, $F(15, 1746) = 21.04$, $p < .001$, $\eta_p^2 = .15$, as was the three-way interaction term, $F(3, 1746) = 7.38$, $p < .001$, $\eta_p^2 = .01$. Simple effects analyses of this model are summarized in Table 3.2. For participants in the blinded conditions, confidence in prior bias beliefs did not meaningfully influence the impact of these beliefs on quality evaluations; however, in the unblinded conditions, participants with more confidence in their prior bias beliefs tended to be more influenced by those beliefs when making their study quality evaluations. Furthermore, as can be seen by comparing the estimates of participants with similar levels of confidence across blinding conditions, confidence increased the biasing influence of these prior beliefs on unblinded quality evaluations. Thus, participants' confidence in these prior beliefs did not alter their quality evaluations in the blinded conditions, yet this confidence exacerbated the biases observed in the unblinded conditions.

Table 3.2. *Simple effects estimates of prior bias beliefs on study quality evaluations by condition and confidence in prior bias beliefs*

Condition	Confidence	Estimate	SE	95% CI	*β*	*p*
Blinded Con-friendly	Below-average	0.12	0.05	[0.01, 0.23]	0.19	.03
	Average	0.12	0.03	[0.06, 0.19]	0.19	< .001
	Above-average	0.13	0.03	[0.06, 0.20]	0.20	< .001
Blinded Lib-friendly	Below-average	0.13	0.05	[0.04, 0.22]	0.20	.004
	Average	0.13	0.03	[0.07, 0.19]	0.20	< .001
	Above-average	0.13	0.03	[0.07 ,0.20]	0.21	< .001
Unblinded Con-friendly	Below-average	-0.13	0.05	[-0.22, -0.04]	-0.20	.004
	Average	-0.20	0.03	[-0.26, -0.15]	-0.32	< .001
	Above-average	-0.28	0.04	[-0.35, -0.20]	-0.44	< .001
Unblinded Lib-friendly	Below-average	0.16	0.05	[0.06, 0.26]	0.25	.001
	Average	0.29	0.03	[0.23, 0.35]	0.45	< .001
	Above-average	0.42	0.03	[0.35, 0.48]	0.65	< .001

Note: Confidence levels: Below-average = *M - 1SD*, Above-average = *M* + 1*SD*.

Credibility impressions

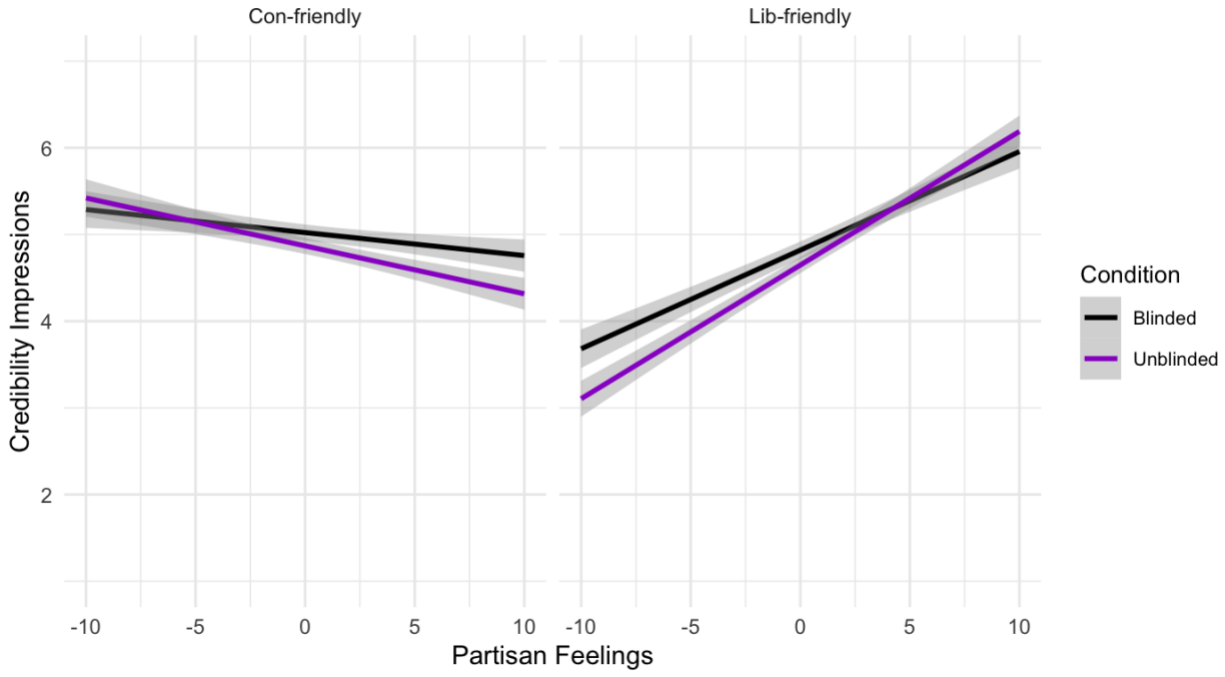
The credibility impressions items had high internal reliability (Cronbach's $\alpha = 0.94$) and were thus averaged into a composite measure, as preregistered. Participants had modestly positive credibility impressions on average ($M = 4.87$, $SD = 1.14$). To assess the influence that participants prior beliefs and the experimental manipulations had on their credibility impressions, in accordance with Research Question 3, I conducted a series of linear regression analyses using the GAMLj package in jamovi.

Influence of partisan feelings

As in the confirmatory analyses of study quality, I first constructed a model predicting study quality evaluations from condition (dummy-coded, 0 = blinded), materials (dummy-coded, 0 = conservative-friendly), partisan feelings (mean-centered), and their interaction. Omnibus tests showed that the overall model was significant, $F(7, 1754) = 76.10$, $p < .001$, $\eta_p^2 = .23$, as was the key three-way interaction term, $F(1, 1754) = 14.48$, $p < .001$, $\eta_p^2 = .01$, illustrated in Figure 3.5. Simple effects analyses showed that partisan feelings were more predictive of credibility impressions in the unblinded conditions. In the conservative-friendly materials, having liberal-leaning feelings predicted forming negative credibility impressions more strongly in the unblinded condition, $b = -0.06$, $SE = 0.01$, $p < .001$, 95% CI [-0.07, -0.04], $\beta = -0.26$, than in the blinded condition, $b = -0.03$, $SE = 0.01$, $p = .004$, 95% CI [-0.04, -0.01], $\beta = -0.12$; in the liberal-friendly materials, having liberal-leaning feelings predicted forming positive credibility impressions more strongly in the unblinded condition, $b = 0.15$, $SE = 0.01$, $p < .001$, 95% CI [0.14, 0.17], $\beta = 0.71$, than in the blinded condition, $b = 0.11$, $SE = 0.01$, $p < .001$, 95% CI [0.10, 0.13], $\beta = 0.53$. The credibility impressions of blinded and unblinded participants within the same materials condition tracked more closely than their study quality evaluations, but the

influence of partisan feelings on credibility impressions was still larger in the unblinded conditions.

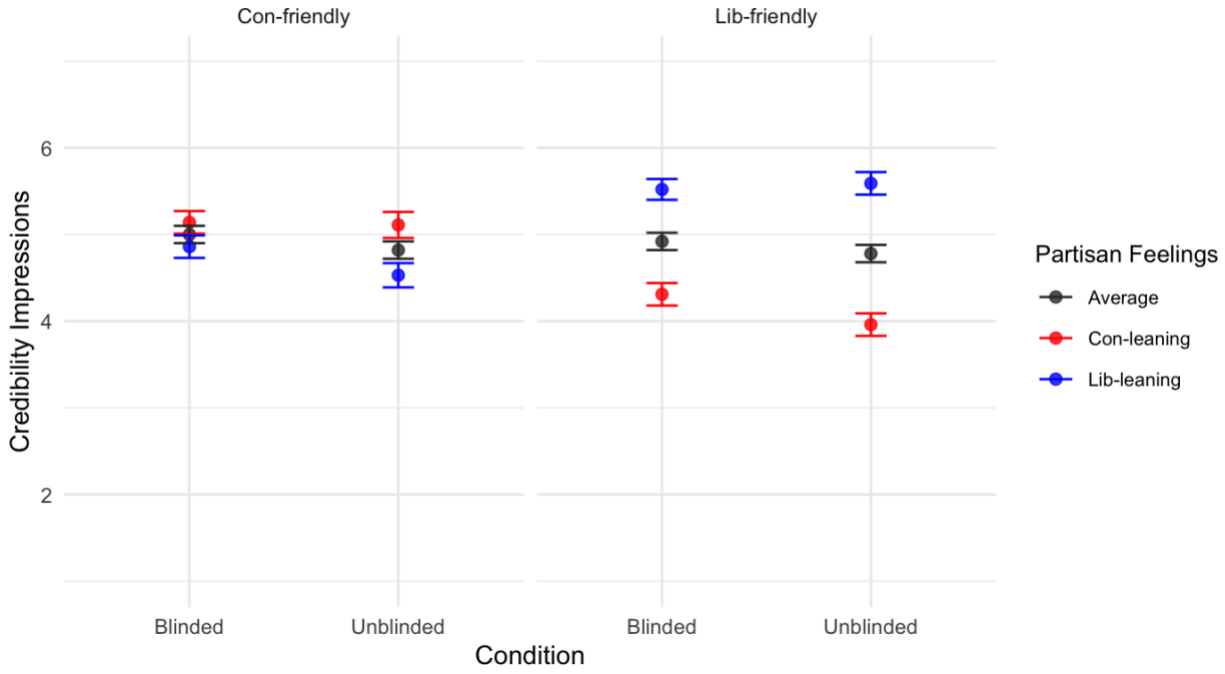
Figure 3.5. Average Credibility Impressions by Partisan Feelings, Materials, and Condition in Study 3
Error bands represent standard errors



This pattern of results is further clarified in Figure 3.6, which depicts the estimated marginal means for this analysis. For example, those with liberal-leaning feelings in the blinded/conservative-friendly condition had significantly more negative credibility impressions of the study than affectively-similar participants who viewed more politically-friendly results in the blinded/liberal-friendly condition. The same pattern held for participants with more conservative-leaning feelings across the blinding conditions. Thus, despite providing statistically equivalent ratings of study quality, participants in the blinded conditions diverged in their credibility impressions of the study after its results were disclosed. Nevertheless, blinding participants quality evaluations did make politically-unfriendly results slightly more palatable to partisans. That is, more liberal-leaning participants had significantly more positive credibility impressions in the blinded/conservative-friendly condition relative to the

unblinded/conservative-friendly condition, and more conservative-leaning participants had significantly more positive credibility impressions in the blinded/liberal-friendly condition relative to the unblinded/liberal-friendly condition. Ultimately, while blinding participants' quality evaluations made them form slightly more positive credibility impressions of politically-unfriendly information, all participants tended to have more positive impressions of a study that produced politically-friendly results.

Figure 3.6. Average Credibility Impressions by Condition, Materials, and Partisan Feelings in Study 3
Error bars represent 95% CIs



To further assess how the blinding manipulation and study quality evaluations influenced credibility impression of the study participants evaluated, and to determine whether these associations differed by partisan feelings, I constructed moderated mediation models using the jAMM package in jamovi. I ran separate models for each set of materials, but the results are substantively identical when using the combined condition variable in a single model. The models included condition (dummy-coded, 0 = blinded) as the predictor, study quality as the mediator, partisan feelings (mean-centered) as the moderator, and credibility impressions as the

outcome variable. These models allowed me to estimate the direct effects that condition had on credibility impressions, the indirect effects that condition had on credibility impressions through evaluations of study quality, and whether these direct and indirect effects differed by partisan feelings. Confidence intervals for the model estimates were calculated using 1000 bootstrap replications.

Table 3.3 provides the full results of the model for the conservative-friendly materials, and Table 3.4 provides the full results of the model for the liberal-friendly materials. In the conservative-friendly materials, there were significant total effects of condition on credibility impressions for participants with average or liberal-leaning feelings, such that being in the unblinded condition was associated with more negative credibility impressions. However, as in the previous studies, these total effects were composed of competing indirect and direct effects. For these participants, there were indirect effects through study quality indicating that being in the unblinded condition caused them to have lower study quality evaluations, which subsequently led them to form more negative credibility impressions. Yet the direct effects for these average and liberal-leaning participants showed that being in the unblinded condition also slightly elevated their credibility impressions. Nevertheless, the indirect effects were larger than the direct effects, so cumulative impact of being in the unblinded condition was that these participants formed more negative credibility impressions. Alternatively, for participants with conservative-leaning feelings, there was a small indirect effect in the opposite direction, such that being in the unblinded condition caused participants to provide more positive study quality evaluations and, subsequently, more positive credibility impressions. This indirect effect was cancelled out by a negative direct effect, however, resulting in a nonsignificant total effect of condition on credibility impressions for these participants.

In the liberal-friendly materials, the mirror opposite pattern of results emerged. There were significant total effects for participants with average and conservative-leaning partisan feelings, and these total effects were driven by indirect effects through quality evaluations. When viewing liberal-friendly materials, those with average and conservative-leaning partisan feelings formed more positive credibility impressions when they made blinded evaluations of the study's quality. Moreover, there was a small indirect effect for participants with liberal-leaning feelings indicating in a small increase in their credibility impressions, yet this effect was negated by a nonsignificant direct effect of condition, yielding a nonsignificant total effect of condition on credibility impressions for these participants. Overall, this pattern of results suggests that, while blinding did not meaningfully increase participants' credibility impression of politically-friendly information, it elevated participants' credibility impressions of politically-unfriendly information by mitigating the influence their partisan feelings had on their quality evaluations.

Table 3.3. Moderated mediation estimates of condition predicting credibility impressions by partisan feelings in the con-friendly materials for Study 3.

Partisan Feelings Group	Type	Effect	Estimate	SE	95% CI	*β*	*p*
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.17	0.07	[0.03, 0.32]	0.09	.02
		Condition ⇒ Study Quality	0.23	0.10	[0.04, 0.43]	0.12	.02
		Study Quality ⇒ Credibility impressions	0.74	0.03	[0.68, 0.79]	0.73	< .001
	Direct	Condition ⇒ Credibility impressions	-0.20	0.06	[-0.32, -0.08]	-0.10	.001
	Total	Condition ⇒ Credibility impressions	-0.03	0.09	[-0.21, 0.15]	-0.02	.75
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.22	0.05	[-0.32, -0.12]	-0.11	< .001
		Condition ⇒ Study Quality	-0.29	0.06	[-0.42, -0.16]	-0.15	< .001
		Study Quality ⇒ Credibility impressions	0.76	0.02	[0.71, 0.81]	0.74	< .001
	Direct	Condition ⇒ Credibility impressions	0.05	0.05	[-0.04, 0.15]	0.03	.25
	Total	Condition ⇒ Credibility impressions	-0.18	0.07	[-0.31, -0.05]	-0.09	.006
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.64	0.07	[-0.78, -0.50]	-0.32	< .001
		Condition ⇒ Study Quality	-0.81	0.08	[-0.98, -0.64]	-0.42	< .001
		Study Quality ⇒ Credibility impressions	0.79	0.04	[0.72, 0.86]	0.75	< .001
	Direct	Condition ⇒ Credibility impressions	0.30	0.07	[0.16, 0.45]	0.15	< .001
	Total	Condition ⇒ Credibility impressions	-0.33	0.09	[-0.51, -0.15]	-0.17	< .001

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 3.4. *Moderated mediation estimates of condition predicting credibility impressions by partisan feelings in the lib-friendly materials for Study 3.*

Partisan Feelings Group	Type	Effect	Estimate	SE	95% CI	β	p
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.52	0.09	[-0.69, -0.34]	-0.20	< .001
		Condition ⇒ Study Quality	-0.60	0.10	[-0.80, -0.39]	-0.28	< .001
		Study Quality ⇒ Credibility impressions	0.86	0.04	[0.79, 0.94]	0.70	< .001
	Direct	Condition ⇒ Credibility impressions	0.17	0.08	[0.02, 0.33]	0.07	.028
	Total	Condition ⇒ Credibility impressions	-0.36	0.10	[-0.55, -0.17]	-0.14	< .001
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.15	0.05	[-0.25, -0.05]	-0.06	.003
		Condition ⇒ Study Quality	-0.19	0.06	[-0.31, -0.06]	-0.09	.003
		Study Quality ⇒ Credibility impressions	0.80	0.03	[0.75, 0.85]	0.67	< .001
	Direct	Condition ⇒ Credibility impressions	0.04	0.05	[-0.05, 0.13]	0.01	.43
	Total	Condition ⇒ Credibility impressions	-0.14	0.07	[-0.28, -0.01]	-0.06	.04
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.16	0.06	[0.05, 0.28]	0.07	.006
		Condition ⇒ Study Quality	0.22	0.08	[0.07, 0.37]	0.11	.005
		Study Quality ⇒ Credibility impressions	0.74	0.04	[0.67, 0.81]	0.64	< .001
	Direct	Condition ⇒ Credibility impressions	-0.10	0.06	[-0.21, 0.01]	-0.04	.073
	Total	Condition ⇒ Credibility impressions	0.07	0.10	[-0.12, 0.26]	0.03	.44

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Influence of prior bias beliefs

The models using prior bias beliefs were substantively identical to the analyses using partisan feelings. These results are presented in full in Appendix C. Blinding participants' quality evaluations made them form slightly more positive credibility impressions of politically-unfriendly information, and this occurred due to the reduced influence that participants' prior bias beliefs had on their study quality evaluations when they evaluated politically-unfriendly results.

Updating partisan feelings and beliefs about partisan bias

Lastly, to further address Research Question 3, I conducted a series of analyses to assess whether participants reported changing their partisan feelings or beliefs about bias after reading the experimental stimuli. In general, participants partisan feelings after reading the stimuli (M = 0.76, SD = 5.30) and their beliefs about partisan bias (M = 4.08, SD = 1.77) were strongly correlated with their initial feelings (r = 0.97) and bias beliefs (r = 0.57). In the subsequent analyses, I examined measures change in partisan feelings (Time 2 feelings – Time 1 feelings) and bias beliefs (Time 2 beliefs – Time 1 beliefs).

Updating partisan feelings

I first constructed a model predicting change in partisan feelings from condition (dummy-coded, 0 = blinded), materials (dummy-coded, 0 = conservative-friendly), partisan feelings (mean-centered), three two-way interactions between these variables, and a three-way interaction. The intercept of this model (b = -0.43, SE = 0.06, 95% CI [-0.55, -0.31]) indicated that participants generally came to feel slightly more positive about conservatives in the blinded/conservative-friendly condition. There was a significant main effect of materials, F(1, 1754) = 63.08, $p < .001$, $\eta_p^2 = .03$, such that viewing the liberal-friendly materials was associated

with participants updating their feelings toward being relatively more positive about liberals, b = 0.66, SE = 0.08, 95% CI [0.50, 0.82], β = 0.51. There was also a main effect of partisan feelings, $F(1, 1754) = 7.09$, $p = .008$, $\eta_p^2 = .00$, such that more liberal-leaning participants came to feel slightly more positive about conservatives on average, b = -0.03, SE = 0.01, 95% CI [-0.05, -0.01], β = -0.13. However, there was not a main effect of condition, $F(1, 1754) = 0.82$, $p = .37$, $\eta_p^2 = .00$, indicating that the blinding manipulation did not directly influence changes in partisan feelings. None of the interaction effects in this model were significant, either (ps ≥ .19). Thus, participants ended the study feeling slightly more positive about political groups whom they read were less biased, and more liberal-leaning participants came to feel more positive about conservatives on average, yet the blinding manipulation did not increase the amount that participants updated their partisan feelings.

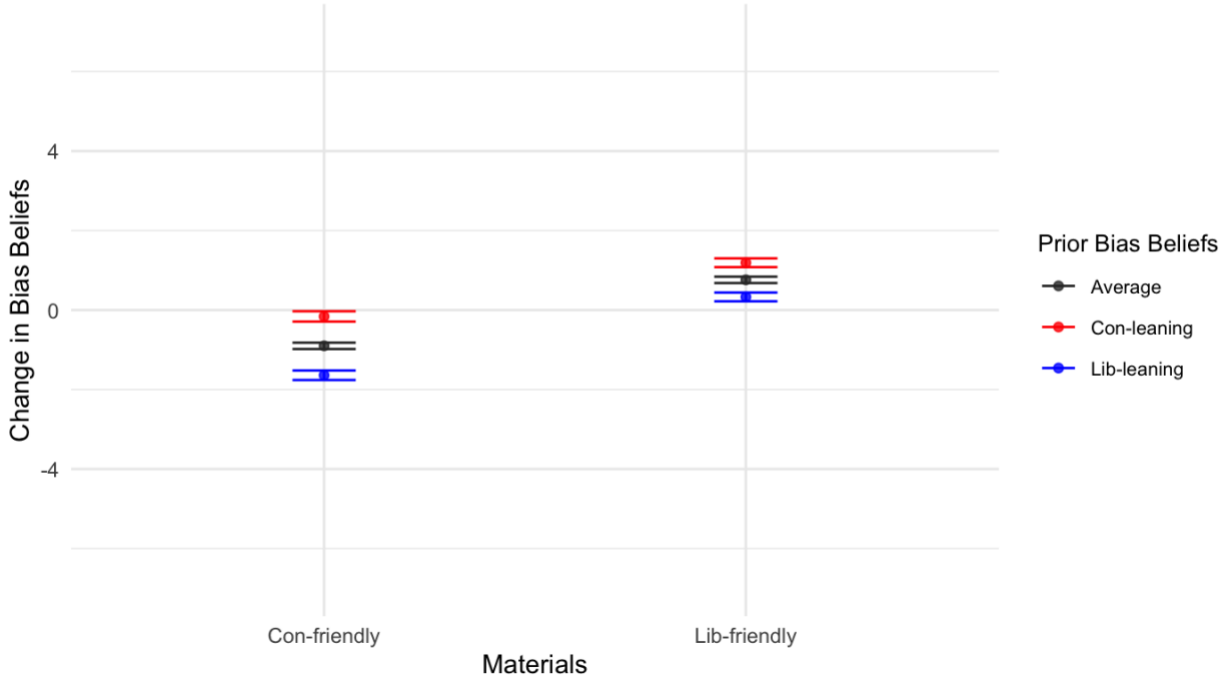
I then created moderated mediation models (one for each set of materials) to assess any indirect effects that condition may have had on participants change in partisan feelings through study quality evaluations and whether that updating process differed across participants with different initial partisan feelings (the results are substantively identical when using combining condition and materials into a single factor variable). These model results are presented in Appendix C. The blinding manipulation yielded significant indirect effects for participants who viewed politically-unfriendly information. That is, being in the blinded conditions caused participants to evaluate the quality of politically-unfriendly information more positively, which in turn increased the amount that participants changed their feelings in response to information that favored their political opponents. Nevertheless, these small indirect effects were nullified by nonsignificant direct effects of condition in the opposite direction, so the total effects of condition on changes in partisan feelings were not significant. Ultimately, blinding participants'

evaluations of politically-unfriendly information made them feel slightly more positive about their opponents compared to co-partisans who made unblinded evaluations of the same information, but these small differences were not significantly impactful when accounting for the total influence that the manipulation had on participants' partisan feelings.

Updating bias beliefs

In the regression model predicting changes in beliefs about partisan bias from condition, materials, and prior beliefs about bias, the intercept (b = -0.91, SE = 0.06, 95% CI [-1.02, -0.80]) indicated that participants came to believe that conservatives are less biased than they originally believed in the blinded/conservative-friendly condition. There was a significant main effect of materials, $F(1, 1754) = 461.74$, $p < .001$, $\eta_p^2 = .21$, and a main effect of prior beliefs, $F(1, 1754) = 178.21$, $p < .001$, $\eta_p^2 = .09$, that was qualified by a significant interaction between prior beliefs and materials, $F(1, 1754) = 17.81$, $p < .001$, $\eta_p^2 = .01$. As illustrated in Figure 3.7, participants with more liberal-leaning prior bias beliefs who read conservative-friendly results updated their beliefs significantly more in the direction of the presented evidence than did those with conservative-leaning priors who read liberal-friendly results. There was not a main effect of condition, $F(1, 1754) = 0.08$, $p = .78$, $\eta_p^2 = .00$, nor any other significant interactions (ps $\geq .36$), indicating that the blinding manipulation did not make participants more open to updating their bias beliefs.

Figure 3.7. Average Change in Bias Beliefs by Materials and Prior Bias Beliefs in Study 3
Error bars represent 95% CIs



The moderated mediation analyses, presented in full in Appendix C, yielded similar results as the comparable analyses using partisan feelings. When evaluating politically-unfriendly information, blinding participants' quality evaluations influenced them to update their bias beliefs in the direction of the presented evidence; nonetheless, the manipulation yielded other countervailing influences on belief updating which, in total, resulted in null effects of blinding on belief-updating.

Overall, participants updated their partisan feelings and beliefs about partisan bias in response to the information they were presented. While participants who made blinded quality evaluations of politically-unfriendly information were influenced to update their beliefs in the direction of the presented evidence slightly more than their unblinded counterparts, the blinding manipulation produced countervailing effects that reduced the condition differences in belief change to nonsignificance.

Discussion

In Study 3, participants were randomly assigned to read about a scientific study which found that either liberals or conservatives are more biased than their political opponents. Participants were also randomly assigned to evaluate the methodological quality of the study either knowing the results (blinded) or after knowing its results (unblinded). Surprisingly, across both sets of materials, participants who made unblinded evaluations were significantly influenced by their partisan feelings and prior beliefs when making quality evaluations, such that more liberal-leaning participants provided more positive quality evaluations than more conservative-leaning participants in the blinded conditions. It is not clear why these effects emerged. Nevertheless, whatever the reasons for the differences in epistemic preferences that were observed in the blinded conditions, the measurement error it reflects should equally apply to the measurement of unblinded participants' prior beliefs. To demonstrate a directional bias in quality evaluations, I did not need to show that *only* unblinded participants were influenced by their partisan feelings and prior beliefs; rather, I needed to determine whether unblinded participants' evaluations were influenced significantly *more* by these beliefs than their blinded counterparts.

As illustrated in Figures 3.1-3.4, I found that unblinded participants were indeed influenced by their prior beliefs significantly more than were blinded participants, indicating that the unblinded participants evaluations were biased by these prior beliefs. Moreover, the observed biases were stronger for participants who viewed politically-unfriendly information. Liberal-leaning participants displayed stronger biases when evaluating the conservative-friendly results, and conservative-leaning participants displayed stronger biases when evaluating the liberal-

friendly results. These findings strongly supported my hypothesis regarding Research Question 1 and were consistent with the results of my three previous studies.

The results of Study 3 also supported my hypothesis for Research Question 2, for unblinded participants' partisan feelings and prior bias beliefs both exerted directional, biasing influences on their quality evaluations. Although these separate prior belief measures exerted independent effects on participants' evaluations, most of their biasing influence was accounted for by their shared influence. As in Study 2, this suggested that "hot" feelings and "cold" expectations were intertwined in biasing unblinded partisans' evaluations. Moreover, the strength of the observed biases did not significantly differ between more liberal and more conservative participants. Study 3 thus provided evidence consistent with accounts of symmetrical partisan biases, and it showed that, as in my previous studies, these biases appear to be stronger when partisans evaluate politically-unfriendly information.

Additionally, to further address Research Question 2, I tested whether various individual difference measures moderated any the observed biases. I found that participants who were more confident in their prior beliefs about bias displayed stronger evaluative biases, as did partisans who more strongly identified with a political party. However, unlike in Study 2, I did not find evidence that analytic thinking exacerbated evaluative biases, and intellectual humility also did not significantly intensify or mitigate the biasing influences of partisans' prior beliefs on their unblinded evaluations. In aggregate, these results do not paint a clear picture of the drivers of partisan bias, a point I will return to in the General Discussion.

Lastly, like in Studies 1a and 2, blinding participants' evaluations of politically-unfriendly information made them form more positive credibility impressions of the study. This indicated that blinded participants found politically-unfriendly information to be more credible,

believable, and trustworthy than their unblinded counterparts. However, while there were also indications that the blinding manipulation influenced participants to update their beliefs about partisan bias to be more in line with the presented evidence, the manipulation also yielded countervailing effects that mitigated the amount of belief updating that blinded participants undertook. It is not clear why the blinding manipulation induced these counteracting effects on belief updating, yet similar effects emerged across all three of the previous studies as well. Despite inducing participants to form more positive quality evaluations and credibility impressions of politically-unfriendly information, the blinding manipulation did not significantly increase how much participants ultimately assimilated the new information into their beliefs.

GENERAL DISCUSSION

Across four preregistered studies, using multiple topics and measuring different beliefs, I compared evaluations of a study's quality between participants who either made *blinded* evaluations—judging the study's methodological merits before knowing the results—or unblinded evaluations of the same information. This experimental design allowed me to test two competing accounts of partisan information processing more strictly than has been done in prior research. The null hypothesis, based on rationalist accounts of information processing and belief updating (Bullock, 2009; Gerber & Green, 1999; Jones & Love, 2011), was that partisans' prior beliefs would not influence their quality evaluations significantly differently in the unblinded conditions than in the blinded conditions. Proponents of rationalist accounts of partisan information processing contend that partisans (or, at least, some partisans) use their prior beliefs in an accuracy-motivated fashion; if this is true, then partisans' beliefs should not influence their methodological evaluations significantly differently regardless of whether they know the results such methods produced. This is because the political-friendliness of new information does not actually reflect its quality to a purely rational evaluator (Kelly, 2008; Ballantyne, 2019; Clark & Winegard, 2020; Carter & McKenna, 2020; Druckman & McGrath, 2019). The alternative hypothesis, which I derived from motivational accounts of partisan information processing (Ditto et al., 2019a; Kahan, 2016), was that partisans' prior beliefs would influence their quality evaluations significantly differently across conditions, indicating a directional bias in the unblinded evaluations. Critically, my experimental design allowed for comparison of co-partisans' evaluations across conditions, and "[h]olding tastes constant" (Gerber & Green, 1999, pg. 206) in this fashion enabled stronger tests of the focal hypotheses (Baron & Jost, 2019; Druckman & McGrath, 2019, Tappin et al., 2020c).

In every study, I found evidence that disconfirmed the null hypothesis of Research Question 1 and lent support to my alternative hypothesis. Compared to blinded individuals who held equivalent prior beliefs, partisans who made unblinded evaluations of the same information were influenced significantly more by their prior beliefs when making those evaluations. Partisans' prior beliefs were inhibited from unduly influencing their quality evaluations when they did not know the results, yet these beliefs exerted directional influences on partisans who knew the results when making their evaluations. Indeed, by modeling how participants' prior beliefs predicted their quality evaluations between conditions, I was able to quantify the average directional influence these beliefs exerted as the difference between the slopes of unblinded and blinded evaluations. Moreover, the effects I observed were primarily driven by unblinded partisans providing negatively biased evaluations of politically-unfriendly information than their blinded counterparts. The difference between the estimated marginal means of co-partisans' unblinded and blinded evaluations quantified the average level of bias demonstrated by these groups of participants, and these differences were consistently larger for groups of participants who evaluated politically-unfriendly information. Overall, by developing an experimental paradigm to address the major limitations of past research, I provided the clearest evidence to date that partisans can be biased by a broad range of prior beliefs when they evaluate scientific information of political relevance.

As reflected in Research Question 2, I was also interested in whether certain types of beliefs were the drivers of partisan bias. In each study, I measured beliefs considered to be more reflective of "hot," affective processes (prior support in Studies 1a-2 and partisan feelings in Study 3) and beliefs theorized to be the product of more "cold," cognitive processes (prior efficacy beliefs in Studies 1a-2 and prior bias beliefs in Study 3). While there was variation

between studies, the results of my most highly-powered analyses (Studies 2 and 3) indicated that these types of beliefs exerted independent, yet highly overlapping, influences on partisan evaluations. Moreover, I also tested whether various individual difference measures exacerbated or mitigated the observed partisan biases, and the results of these analyses did not yield clear theoretical interpretations. As I will discuss more in the following section, there were effects that seemed to reflect cognitively-driven biases and others that appeared to reflect affectively-driven biases. In total, there was little indication that either affective or cognitive processes were the fundamental driver of partisan biases; rather, my results suggested that cognitive and affective processes are deeply intertwined in the emergence of partisan bias.

Additionally, throughout my studies, I tested how the blinding manipulation influenced participants' downstream judgments of how credible they found presented study and how much they updated their beliefs after considering that new information. I did not have any strong hypothesis for these analyses, but I conducted analyses to address Research Question 3 and explore the effects of blinding on how partisans' ultimately processed the information they evaluated. Across studies, the results indicated that blinded participants who evaluated politically-unfriendly information formed significantly more positive credibility impressions and updated their beliefs slightly (but, except for Study 2, nonsignificantly) more than their unblinded counterparts. Moderated mediation analyses showed that the condition differences in credibility impressions and belief change for these participants could be explained by the blinded participants' greater quality evaluations. In other words, prohibiting participants from offering biased evaluations of politically-unfriendly information increased how believable they found the results and, to a lesser extent, increased the degree to which they subsequently updated their beliefs compared to their unblinded counterparts. However, while blinding indirectly increased

participants' credibility impressions and belief updating of politically-unfriendly information by elevating their quality evaluations, blinding also produced countervailing direct effects. Consequently, despite the significant indirect effects, blinded participants generally did not have significantly more positive credibility impressions and belief updating after evaluating politically-unfriendly information. It is not clear why the manipulation yielded these contrasting effects, though as I return to in the upcoming section on practical implications, it will be imperative to gain a deeper understanding of this suite of effects before blinding should be considered as an intervention for attitude change. Nevertheless, I demonstrated that partisans' biased evaluations can influence how they assimilate new information into their beliefs, which can inform descriptive models of partisan information processing (Druckman & McGrath, 2019; Gerber & Green, 1999; Kahan, 2016; Tappin et al., 2020c), as well as more general models of learning and belief updating (Jones & Love, 2011; Uhlmann, 2011).

Theoretical Implications

Despite decades of research claiming to document the effects of directional motivations on judgment, claims of bias in political science and social psychology have often been met with "post hoc rationalism" (Uhlmann, 2011, pg. 214), such as Bayesian models of belief updating (Baron & Jost, 2019). Invocations of accuracy-motivated accounts for seemingly obvious directional biases are not without merit, as there have been notable limitations in past research that has sought to document partisan bias (Ditto et al., 2019b, Druckman & McGrath, 2019; Tappin et al., 2020c). Indeed, descriptively disentangling accuracy-motivated from directional-motivated partisan judgments is but the latest evolution of longstanding debates over observational equivalence in social cognition (Bullock, 2009; Ditto, 2009; Erdeyli, 1974; Gerber & Green, 1999; Miller & Ross, 1975; Tetlock & Levi, 1982). Nevertheless, such Bayesian just-

so stories, and post hoc rationalizations more broadly (Uhlmann, 2011), have functionally relegated motivational accounts of information processing to a second-class theoretical status, an alternative juxtaposed against the normative, rationalist, accuracy-motivated default (Ditto, 2009). In these experiments, I sought to embrace this dynamic by putting the rationalist, purely-accuracy motivated account of partisan judgment to the test, setting it as the null hypothesis against which dispositive evidence may or may not emerge.

Ultimately, the presented results clearly disconfirmed strong accounts of political reasoning as a purely accuracy-motivated process. As I hypothesized, partisans' unblinded evaluations were influenced by their prior feelings and beliefs significantly more than their blinded counterparts, demonstrating a directional influence on their evaluations. However, despite the misconceptions of many proponents of rationalist accounts of political reasoning, my results do not disconfirm Bayesian models of belief updating. As stated most clearly by Druckman and McGrath, one can be both accuracy- and directionally-motivated and still update beliefs in a manner consistent with Bayesian logic (2019, pg. 112). Bayesian logic emphasizes the importance of people's prior beliefs in how they come to understand the world, yet Bayesian models are often underspecified regarding where, and to what extent, prior beliefs exert their influence in the information processing stream (Jones & Love, 2011). Including a path through which prior beliefs can bias evaluations of new information is thus not inherently inconsistent with adopting a Bayesian framework of belief updating. In other words, people can be both Bayesian and biased. Conflating rationalist theories of reasoning with Bayesian models of reasoning has led to unnecessary confusion in discussions of bias, and such conflation has been enabled by the lack of specificity in most Bayesian models of human cognition. Demonstrating that partisans exhibit directional biases, as I have done in these studies, provides stronger

empirical support to the contention that—to be descriptively accurate—Bayesian models must account for evaluative biases (Uhlmann, 2011).

Rather than pitting accuracy-motivated and directionally-motivated accounts of reasoning against one another, future research should seek to specify the extent to which these different motivations influence various parts of the information processing stream in different contexts. So much energy has been spent litigating whether directional motivations exist (Ditto, 2009) that we do not currently understand the degree to which accuracy and directional motives inform one another or operate independently. For instance, the present results indicated that participants with weaker political motivations (as indexed by weaker prior beliefs and partisan affiliations) exhibited weaker biases, yet it is not clear whether these participants were merely less directionally-motivated or were also more accuracy-motivated than their more partisan peers. Acknowledging that both directional and accuracy motivations can influence reasoning may help political psychologists connect with other fields, such as clinical psychology, to develop theories that can explain more diverse psychological phenomena. For example, although they may initially seem like unrelated topics, partisans biases and placebo effects may both be driven by how much people want and expect a particular outcome (Liu, 2022). Likewise, identifying how to help mentally ill individuals update their delusional beliefs to be in greater accordance with reality (Kube & Rozenkrantz, 2021) may shed light on ways to foster political consensus between polarized partisans who perceive disjunctive realities. On January 6th, 2021, people who believe that an organization of Satanic, cannibalistic pedophiles conspired against former President Donald Trump (Kunzelman, 2020) tried to overthrow the lawful results of a presidential election; one could be forgiven for mistaking this political "movement" for a mental health crisis. Discarding extreme rationalist accounts of partisan judgment, as is justified by the

present results, will facilitate more constructive theoretical debates that may yield clearer descriptions of political reasoning and human psychology.

A limitation of the present research was that it did not clearly illuminate the mechanisms through which political beliefs bias partisans' unblinded quality evaluations. While the results related to Research Question 2 suggested that there were both cognitive and affective drivers of partisan bias, few of the individual difference measures significantly moderated the observed biases. In Study 2, participants' positive and negative affect mediated a statistically significant, but practically insubstantial, amount of the variance between blinded and unblinded co-partisans' quality evaluations, suggesting that such affective reactions were only a small part of the observed biases. Similarly, there were indications that moral conviction (a more affective individual difference; Skitka & Wisneski, 2011) may have exacerbated the biases observed in Study 2, but these effects were too small to be reliably estimated. However, the moderation analyses I ran throughout my studies did not clearly favor more cognitive accounts of partisan bias, either. For instance, analytic thinking (a more cognitive individual difference; Frederick, 2005) seemed to slightly exacerbate partisan biases in Study 2, but these effects were not replicated in Study 3. The clearest evidence of moderation was observed in Study 3, where participants' with greater confidence in their prior beliefs and stronger partisan identifications exhibited stronger biases. Yet these clearer statistical results do lend themselves to obvious theoretical interpretations: confidence in one's beliefs and strength of partisan identification are likely both subject to affective and cognitive precursors. It is tempting to suggest that future research may be able to clarify the relative influences of "hot" desires and "cold" expectations in political reasoning. Nevertheless, if we accept that partisans' desires and expectation are often deeply intertwined, then it may be more fruitful for political psychologists to investigate how and

why such coherence emerges (e.g., Liu & Ditto, 2014) rather than seeking to artificially disambiguate their shared influence.

One may reasonably wonder whether my results, derived from online samples of Americans on a few topics, are generalizable within or beyond the American context. There are at least two ways in which one can think about the generalizability of the present findings. In a narrower sense, it is likely that there will be substantial heterogeneity in the magnitude of directional biases observed across samples and political topics (Yarkoni, 2022). Indeed, even in these four studies, which all sampled from the same online population, participants exhibited weaker biases when evaluating a study about trigger warnings (Study 1b) than when evaluating studies about capital punishment and partisan bias (Studies 1a, 2, and 3). This should inspire caution when discussing the implications of these studies, particularly Study 3, in debates over symmetries and asymmetries in partisan bias (Baron & Jost, 2019, Ditto et al., 2019a, 2019b). The presented results are more consistent with a symmetrical account of partisan bias, for liberals and conservatives did not meaningfully differ in the amount of bias exhibited in Study 3. That said, the results of future studies will likely vary as a function of the political topics or groups being evaluated (e.g., for a review of asymmetries in political prejudice toward various groups, see Crawford & Brandt, 2020). Although some have posited specific personality or cognitive differences between partisans, such as conservatives' greater need for epistemic certainty, that may make conservatives more prone to bias (Jost, et al., 2007), the relationship between conservatism and the need for certainty varies substantially across cultures (Federico & Malka, 2018), and American conservatives' greater needs for certainty do not meaningfully predict partisan biases (Guay & Johnston, 2021). While I am agnostic as to whether partisan biases are symmetrical in aggregate, future research will benefit from focusing more narrowly on

(a)symmetries in bias in relation to specific topics, examining differences in personality traits and social dynamics that may account for such (a)symmetries in a particular context (e.g., for asymmetries in democratic attitudes, see Benjamin et al., 2021). Beyond the focus on symmetries and asymmetries, more research is needed to replicate and extend the present findings before we can claim to know the general prevalence or magnitude of partisans' evaluative biases.

Nevertheless, there is a broader sense in which the present results should be broadly generalizable. Unlike much research in psychology (Klayman & Ha, 1987; Uchino et al., 2010), these studies were designed to provide disconfirmatory evidence against a theory. By providing clear evidence against a purely accuracy-motivated account of partisan evaluations in one population, we can dismiss this theory in *any* population that operates through similar cognitive and social architecture. Unless one identifies problems with the internal validity of the present experiments or can offer a mechanism to account for a particular population's immunity to directional influences, then we should assume that other populations are at least comparably prone to biases until proven otherwise. In other words, the disconfirmatory evidence provided by these studies shifts the burden of proof onto theorists of political reasoning who claim that partisans are purely accuracy-motivated. Such proponents must now show that directional motivations yield no influence on their population's evaluations to justify their theory.

Practical Implications

The present research highlights the potential of blinding techniques to reduce evaluative biases in contexts where evaluators may be swayed by directional motivations, political or otherwise. For instance, the benefits of blinding have been clearly demonstrated in employment contexts, where blinding evaluations of job candidates has been shown to reduce gender biases in hiring decisions (Goldin & Rouse, 2000; Uhlmann & Cohen, 2005). Interestingly, the original

research in this area found gender biases against women, but more recent replications have found biases against men in unblinded hiring evaluations (Tierney et al., 2020). This development illustrates one of the great utilities of blinding: one need not know the direction of a bias for blinding to be an effective bias prevention strategy. Whatever twists and turns may emerge within a sociopolitical zeitgeist, blinding strategies can help people avoid the "mental contamination" (Wilson & Brekke, 1994) of irrelevant decision criteria. However, while a recent study found that people think they should blind themselves to potentially biasing information, very few of these people chose to actually blind themselves when the opportunity was available to them (Fath et al., 2022). Indeed, people who perceive themselves to be more objective tend to be more biased than those who think themselves to be less objective (Uhlmann & Cohen, 2007), and people who think they are impartial may be the least likely to pursue blinding strategies. The discrepancies between perceived and actual impartiality demand further investigation (e.g., Pronin, 2007); nevertheless, for the time being, such findings suggest that successful implementation of blinding strategies will require collective, structural, top-down efforts rather than a reliance on individual responsibility. This may be especially true in the communication of politically-relevant science in our highly polarized era (Druckman, 2017; Finkel et al., 2020), as political partisans typically perceive themselves to be more objective than their opponents (Ditto et al., 2019a). Collaborations between social media platforms, journalists, and scientists may facilitate the development of communication strategies that can mitigate biased evaluations of politically-relevant research.

Nevertheless, more basic research will first be required to map the downstream influences that blinding evaluations can have on the belief updating process. In my studies, blinding politically-unfriendly information had countervailing effects on participants' credibility

impressions and belief change; although blinding increased quality evaluations and thus, indirectly, increased credibility impressions and belief updating in the direction of the presented evidence, blinding also exerted a direct influence on these outcomes that restricted such positive impression formation and belief change. I do not know why the direct, restricting influences emerged, but there are several hypotheses worth testing in future research. First, blinding participants' evaluations may feel unnatural given the way such information is typically evaluated, and ultimately presenting these people with politically-unfriendly results may make them more cognizant of the artificial setting in which they are evaluating information. This may result in participants finding the presented information less believable than they might in a more naturalistic setting. Exploring ways to make evaluative blinding manipulations subtler and more ecologically valid could help reduce the countervailing effects that the manipulations have on downstream judgments. It may also be the case that providing blinded evaluations of politically-unfriendly information induces a sense of self-perceived objectivity, so people may feel licensed to denigrate the ultimate credibility of the research once its results are known to them (Uhlmann & Cohen, 2007). Investigating these and other viable hypotheses will be needed to understand the potential of blinding strategies for foster belief-updating and, ultimately, political consensus. These inquiries may also aid researchers in identifying other forms of bias that can affect the partisan information-processing stream. While blinding reduced partisans' evaluative biases, additional interventions will likely be required to foster political consensus on the myriad issues that rely on appeals to scientific evidence.

Finally, while the focus of this research has been on lay evaluations of science, future research should investigate the extent to which partisan biases may influence scientists' evaluations of research—particularly in the social and behavioral sciences. Researchers do not

often agree on the quality of research even in the absence of politics; for example, evaluations of NIH grant proposals are better predicted by the idiosyncratic preferences of reviewers than by the number of weakness identified by those reviewers in a particular proposal (Pier et al., 2018). Given this highly subjective peer-review process, it is plausible that scientists evaluate research that runs counter to their cherished theoretical or political commitments more skeptically than research that affirms those commitments. Evaluative competence and domain-area expertise may mitigate partisan biases, yet such expertise may instill self-perceptions of objectivity that, as has been shown in lay evaluations (Uhlmann & Cohen, 2007), could undermine experts' evaluations of politically-unfriendly results. The political slant of published research in psychology has not been found to relate to its replicability (Reinero et al., 2020), but no research to date has examined how scientists' political beliefs may influence their methodological evaluations of politically-relevant research. Biases may emerge in the evaluation of journal submissions, distorting the landscape of research that is ultimately published on a particular topic. Indeed, if we take them at their word, some researchers find it "reasonable and appropriate for people to ask whether the study's conclusions agree with their preexisting beliefs" when evaluating its methodological quality (Baron & Jost, 2019, pg. 297). As others have noted, "It would be remarkable indeed if scientists were immune to the empirical phenomena we study" (Uhlmann, 2011, pg. 214).

Reforms to scientific practices in the wake of psychology's replication crisis, such as the emergence of registered reports, have been promoted to reduce publication biases (Nosek & Lakens, 2014; Chambers et al., 2015; Munafò et al., 2017). Broader adoption of such practices can help scientists mitigate partisan and other directional biases that may underlie publication biases in some fields. Additional reforms, such as blinding analyses (MacCoun & Perlmutter,

2015) and creating blinded peer-review processes for non-registered reports—for example, creating tiered reviews where reviewers are blinded to a study's results when they first evaluate its methods—may also help reduce biases. While other reforms, such as increasing the political diversity of researchers (Duarte et al., 2015), would be beneficial in other ways, I believe that reducing evaluative biases is a more tractable and sustainable path of quality control than intentionally introducing countervailing biases. Maintaining the appropriate balance of partisans within a field (e.g., determining what levels of imbalance may be considered appropriate) will not be as feasible as leaning further into blinding strategies that are already the conceptual basis of successful scientific reforms. Crucially, and as mentioned earlier in this section, blinding techniques are equal-opportunity bias prevention: blinding strategies can help reduce conservative biases in one journal or field while similar procedures can help reduce prevailing liberal biases in a different subdiscipline. Given the costs such reforms likely require, it may be worth investigating the extent to which directional biases influence scientists' evaluations before pursuing reforms in earnest. Yet if we are not certain that scientists are immune to such evaluative biases—or if we are not convinced that our successors will be purely objective—then it may be worth investing in such reforms even before we know the prevalence of scientists' biases.

Conclusion

What people believe and want to be true can inhibit them from perceiving what is actually true. As illustrated by the quote that opened this dissertation, this sentiment is prevalent in folk epistemology. Yet psychologists have been adept at constructing rationalizations to explain purportedly biased behaviors, making it difficult to determine whether motivated biases even exist. While it remains challenging to identify whether any singular evaluation is biased, I

demonstrated that, on average, partisans are directionally biased by their prior feelings and beliefs when making unblinded evaluations of politically-relevant science. Blinding partisans' quality evaluations revealed and reduced these biases, particularly in evaluations of politically-unfriendly information. The dispositive evidence I provided against accounts of purely rational, accuracy-motivated partisan judgment signals that such accounts need to be put to rest. Focusing empirical and theoretical attention on *when*, *how*, and *to what degree* people exhibit biases will be more fruitful than debating *if* such biases exist. Further documenting the extent to which lay and expert evaluations are subject to directional biases will aid scientists in developing reforms to conduct and communicate their research more effectively.

REFERNCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York, NY: Harper.

Areni, C. S., & Lutz, R. J. (1988). The Role of Argument Quality in the Elaboration Likelihood Model. *Advances in Consumer Research Volume*, *15*, 197–203. https://doi.org/10.1017/CBO9781107415324.004

Ballantyne, N. (2019). *Knowing our limits*. Oxford University Press.

Bankert, A., Huddy, L., & Rosema, M. (2017). Measuring Partisanship as a Social Identity in Multi-Party Systems. *Political Behavior*, *39*(1), 103–132. https://doi.org/10.1007/s11109-016-9349-5

Baron, J., & Jost, J. T. (2019). False Equivalence: Are Liberals and Conservatives in the United States Equally Biased? *Perspectives on Psychological Science*, *14*(2), 292–303. https://doi.org/10.1177/1745691618788876

Bastardi, A., Uhlmann, E. L., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science*, *22*(6), 731–732. https://doi.org/10.1177/0956797611406447

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science*, *17*(1), 78-98.

Benjamin, R., Laurin, K., & Chiang, M. (2021). Who would mourn democracy? Liberals might, but it depends on who's in charge. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspa0000291

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior*, *36*(2), 235–262. https://doi.org/10.1007/s11109-013-9238-0

Bowes, S. M., Costello, T. H., Lee, C., McElroy-Heltzel, S., Davis, D. E., & Lilienfeld, S. O. (2022). Stepping Outside the Echo Chamber: Is Intellectual Humility Associated with Less Political Myside Bias?. *Personality and Social Psychology Bulletin*, *48*(1), 150-164.

Bullock, J. G. (2009). Partisan bias and the Bayesian ideal in the study of public opinion. *Journal of Politics*, *71*(3), 1109–1124. https://doi.org/10.1017/S0022381609090914

Carter, J., & McKenna, R. (2020). Skepticism motivated: on the skeptical import of motivated reasoning. *Canadian Journal of Philosophy*.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, 1–2. https://doi.org/10.1016/j.cortex.2015.03.022

Clark, C. J., & Winegard, B. M. (2020). Tribalism in War and Peace: The Nature and Evolution of Ideological Epistemology and Its Significance for Modern Social Science. *Psychological Inquiry*, *31*(1), 1–22. https://doi.org/10.1080/1047840X.2020.1721233

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics*, *2*(4), 2053168015622072.

Cohen, G. L. (2003). Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Crawford, J. T., & Brandt, M. J. (2020). Ideological (a)symmetries in prejudice and intergroup bias. *Current Opinion in Behavioral Sciences, 34*, 40-45.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-Test instead of student's t-Test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Ditto, P. H. (2009). Passion, Reason, and Necessity: A Quantity-of-Processing View of Motivated Reasoning. In *Delusion, Self-deception, and Affective Influences on Belief Formation* (pp. 23–54). https://doi.org/10.4324/9780203838044

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., … Zinger, J. F. (2019a). At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291. https://doi.org/10.1177/1745691617746796

Ditto, P. H., Clark, C. J., Liu, B. S., Wojcik, S. P., Chen, E. E., Grady, R. H., … Zinger, J. F. (2019b). Partisan Bias and Its Discontents. *Perspectives on Psychological Science, 14*(2), 304–316. https://doi.org/10.1177/1745691618817753

Ditto, P. H., & Lopez, D. L. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63, 568–584.

Ditto, P. H., Munro, G. D., Lockhart, L. K., Scepansky, J. A., & Apanovitch, A. M. (1998). Motivated Sensitivity to Preference-Inconsistent Information. *Journal of Personality and Social Psychology, 75*(1), 53–69. https://doi.org/10.1037/0022-3514.75.1.53

Druckman, J. N. (2017). The crisis of politicization within and beyond science. *Nature Human Behaviour*, *1*(9), 615–617. https://doi.org/10.1038/s41562-017-0183-5

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119. https://doi.org/10.1038/s41558-018-0360-1

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences*, *38*, e130. doi:10.1017/ S0140525X14000430

Ebersole, C. R. (2019). Pre-commitment and Updating Beliefs. Unpublished dissertation.

Erdeyli, M. H. (1974). A new look at the New Look: Perceptual defense and vigilance. *Psychological Review*, *81*, 1–25.

Fath, S., Larrick, R. P., & Soll, J. B. (2022). Blinding curiosity: Exploring preferences for "blinding" one's own judgment. *Organizational Behavior and Human Decision Processes*, *170*, 104135.

Federico, C. M., & Malka, A. (2018). The Contingent, Contextual Nature of the Relationship Between Needs for Security and Certainty and Political Preferences: Evidence and Implications. *Advances in Political Psychology*, *39*, 3–48. https://doi.org/10.1111/pops.12477

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533-536.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Gerber, A., & Green, D. (1999). Misperceptions About Perceptual Bias. *Annual Review of Political Science*, *2*(1), 189–210. https://doi.org/10.1146/annurev.polisci.2.1.189

Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality. *The American Economic Review*, *90*(4), 715–741. https://doi.org/10.4324/9780429494468-41

Guay, B., & Johnston, C. (2021). Ideological Asymmetries and the Determinants of Politically Motivated Reasoning. *American Journal of Political Science*, 1–60.

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People Claim Objectivity After Knowingly Using Biased Strategies. *Personality and Social Psychology Bulletin*, *40*(6), 691–699. https://doi.org/10.1177/0146167214523476

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134

Jost, J. T., Napier, J. L., Thorisdottir, H., Gosling, S. D., Palfai, T. P., & Ostafin, B. (2007). Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity?. *Personality and social psychology bulletin*, *33*(7), 989-1007.

Kahan, D. M. (2016). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 1–16. https://doi.org/10.1002/9781118900772.etrds0417

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self- government. *Behavioral Public Policy*, *1*, 54-86.

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist*. https://doi.org/10.1037/0003-066X.58.9.697

Kelly, T. (2008). Disagreement, Dogmatism, and Belief Polarization. *The Journal of Philosophy*, *105*(10), 611–633.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037//0033-295x.94.2.211

Kruglanski, A. W., Dechesne, M., Orehek, E., & Pierro, A. (2009). Three decades of lay epistemics: The why, how, and who of knowledge formation. *European Review of Social Psychology*, *20*, 146–191. https://doi.org/10.1080/10463280902860037

Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: an integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science*, *16*(2), 247-274.

Kubin, E., Puryear, C., Schein, C., & Gray, K. (2021). Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, *118*(6).

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kunzelman, M. (2020, February 9). 'QAnon' conspiracy theory creeps into mainstream politics. *Associated Press.* https://web.archive.org/web/20210819200201/https://apnews.com/article/donald-trump-us-news-ap-top-news-wisconsin-racine-e230131513bf3df60c76bb1151bc6b7c

Leary, M. R., Diebels, K. J., Davisson, E. K., Isherwood, J. C., Jongman-Sereno, K. P., Raimi, K. T., … Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality & Social Psychology Bulletin*. https://doi.org/10.1177/0146167217697695

Leeper, T. J., & Slothuus, R. (2014). Political parties, motivated reasoning, and public opinion formation. *Political Psychology, 35*, 129–156. https://doi.org/10.1111/pops.12164\

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The Demographic and Political Composition of Mechanical Turk Samples. *SAGE Open*, *6*(1). https://doi.org/10.1177/2158244016636433

Liu, B. S. (2014). The expertise paradox: Examining the role of different aspects of expertise in biased evaluation of scientific information. Unpublished dissertation.

Liu, B. S., & Ditto, P. H. (2013). What Dilemma? Moral Evaluation Shapes Factual Belief. *Social Psychological and Personality Science*, *4*(3), 316–323. https://doi.org/10.1177/1948550612456045

Liu, T. (2022). Placebo Effects: A New Theory. *Clinical Psychological Science, 10*(1), 27-40.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189. https://doi.org/10.1038/526187a

Miller, A. G., McHoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, *64*(4), 561–574. https://doi.org/10.1037//0022-3514.64.4.561

Miller, D. T., & Ross, M. (1975). Self-serving biases in attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*, 213–225.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636–653. https://doi.org/10.1177/0146167297236007

Munro, G. D., Ditto, P. H., Lockhart, L. K., Fagerlin, A., Gready, M., & Peterson, E. (2002). Biased Assimilation of Sociopolitical Arguments: Evaluating the 1996 U.S. Presidential Debate. *Basic and Applied Social Psychology*, *24*(1), 15–26. https://doi.org/10.1207/s15324834basp2401_2

Munro, G. D., & Munro, C. A. (2014). "Soft" Versus "Hard" Psychological Science: Biased Evaluations of Scientific Evidence That Threatens or Supports a Strongly Held Political Identity. *Basic and Applied Social Psychology*, *36*(6), 533–543. https://doi.org/10.1080/01973533.2014.960080

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

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

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., ... & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 2952-2957.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, *11*(1), 37–43. https://doi.org/10.1016/j.tics.2006.11.001

Reinero, D. A., Wills, J. A., Brady, W. J., Mende-Siedlecki, P., Crawford, J. T., & Van Bavel, J. J. (2020). Is the Political Slant of Psychology Research Related to Scientific Replicability? *Perspectives on Psychological Science*, (2018). https://doi.org/10.1177/1745691620924463

Scurich, N., & Shniderman, A. (2014). The selective allure of neuroscientific explanations. *PLoS ONE*, *9*(9), 1–6. https://doi.org/10.1371/journal.pone.0107529

Seybold, M. (2016). *The apocryphal Twain: "Things we know that just ain't so."* Center for Mark Twain Studies. https://marktwainstudies.com/the-apocryphal-twain-things-we-know-that-just-aint-so/

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, 8, 895–917.

Skitka, L. J., Liu, J. H. F., Yang, Y., Chen, H., Liu, L., & Xu, L. (2012). Exploring the Cross-Cultural Generalizability and Scope of Morally Motivated Intolerance. *Social Psychological and Personality Science*, *4*(3), 324–331. https://doi.org/10.1177/1948550612456404

Skitka, L. J., & Wisneski, D. C. (2011). Moral conviction and emotion. *Emotion Review*, *3*(3), 328–330. https://doi.org/10.1177/1754073911402374

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020a). Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General*, *150*(6), 1095-1114

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020b). Bayesian or biased? Analytic thinking and political belief updating. *Cognition*, *204*, 104375.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020c). Thinking clearly about causal inferences of politically motivated reasoning: why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, *34*, 81–87. https://doi.org/10.1016/j.cobeha.2020.01.003

Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology*, *18*, 68–88.

Thaler, M. (2019). The "Fake News" Effect: An Experiment on Motivated Reasoning and Trust in News. Preprint.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

Tierney, W., Hardy, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., … Uhlmann, E. L. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes*, *161*(July), 291–309. https://doi.org/10.1016/j.obhdp.2020.07.002

Uchino, B. N., Thoman, D., & Byerly, S. (2010). Inference Patterns in Theoretical Social Psychology: Looking Back as We Move Forward. *Social and Personality Psychology Compass*, *4*(6), 417–427. https://doi.org/10.1111/j.1751-9004.2010.00272.x

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*(2), 207–223. https://doi.org/10.1016/j.obhdp.2007.07.001

Uhlmann, E. L. (2011). Post hoc rationalism in science. *Behavioral and Brain Sciences*, *34*(4), 214-214.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063.

Washburn, A. N., & Skitka, L. J. (2017). Science Denial Across the Political Divide. *Social Psychological and Personality Science*, 194855061773150. https://doi.org/10.1177/1948550617731500

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. https://doi.org/10.1002/9781118890233.ch25

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*.

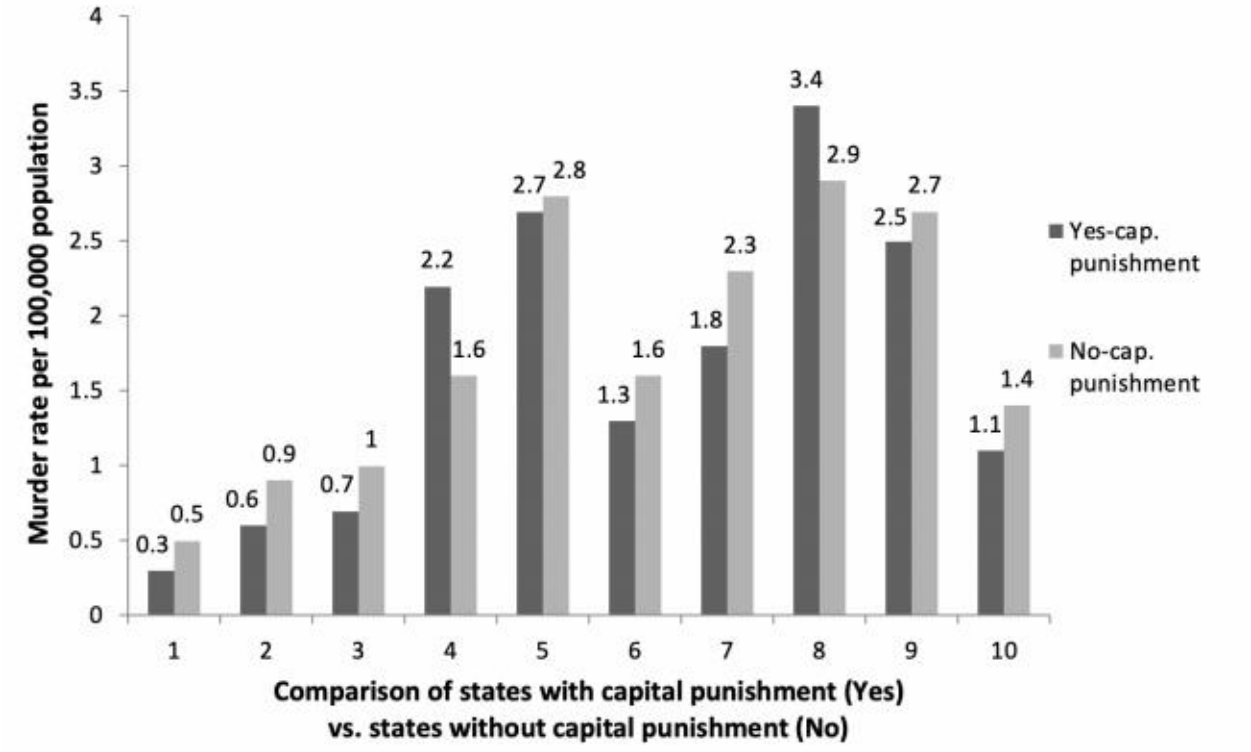
APPENDIX A: Studies 1a and 1b Supplement

Figures accompanying the stimuli in Studies 1a and 1b

Study 1a

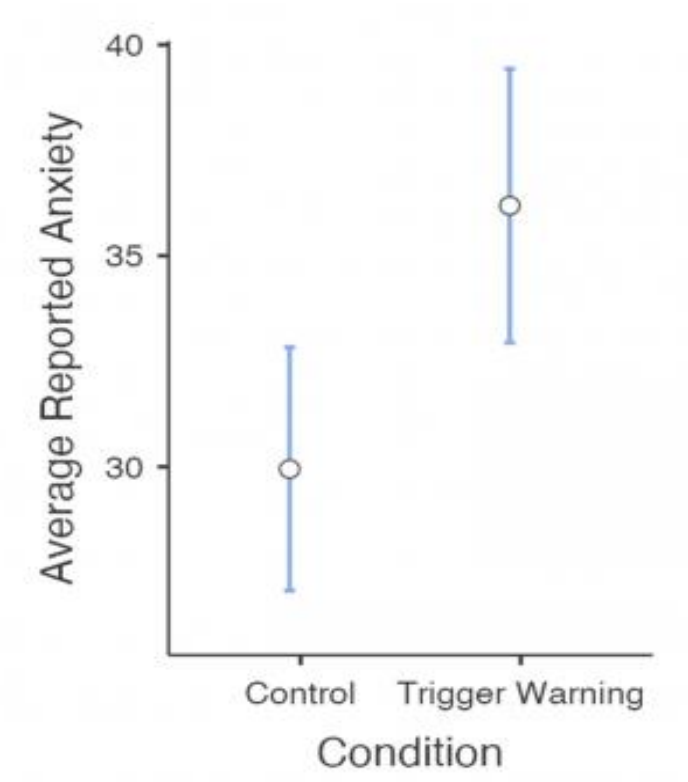
Murder rate in 2006 for States with Capital Punishment (Yes) and States Without Capital Punishment (No)

Pair	State	Murder rate	Has Capital Punishment?	Pair	State	Murder rate	Has Capital Punishment?
1	A	0.3	Yes	6	K	1.3	Yes
	B	0.5	No		L	1.6	No
2	C	0.6	Yes	7	M	1.8	Yes
	D	0.9	No		N	2.3	No
3	E	0.7	Yes	8	O	3.4	Yes
	F	1.0	No		P	2.9	No
4	G	2.2	Yes	9	Q	2.5	Yes
	H	1.6	No		R	2.7	No
5	I	2.7	Yes	10	S	1.1	Yes
	J	2.8	No		T	1.4	No



Study 1b

Average Reported Anxiety to Distressing Passages Across Experimental Conditions (with 95% Confidence Intervals)



Demographic questions

Age

Participants were asked to report their age on a dropdown menu that included age from 18 to 99.

Sex

Participants were asked to select their sex from three response options: “Female,” “Male,” or “Other.” The “Other” option included a text box for participants to describe their sex.

Ethnicity

Participants were asked "What is your race/ethnicity?" and asked to select one of the following options: "African American, Black, African, Caribbean"; "East Asian (Chinese, Korean, Japanese, etc.); "South Asian (Indian, Pakistani, Sri Lankan, etc.)"; "Southeast Asian (Vietnamese, Cambodian, Filipino, etc.)"; "European American, White, Anglo, Caucasian"; "Hispanic American, Latino(a), Chicano(a), Mexican, Columbian"; "Pacific Islander (Micronesian, Melanesian, Samoan, etc.)"; "Native Hawaiian, American Indian, Alaskan Native"; "Biracial, Multiracial - please describe your race below", which included an open text box; "Other - please describe your race below", which included an open text box; and "I decline to answer."

Education

Participants were asked "What is the highest level of education you have completed?" and asked to select one of the following options: "Some high school or less (no HS diploma)," "High school diploma," "Some college (no degree)," Associates degree," "Bachelors degree," "Masters degree," "Professional or doctorate degree (PhD, JD, MD, etc.)."

Household income

Participants were asked "What is your yearly household income?" and asked to select one of 19 income bins ranging from "Less than $5,000" to "$175,000 or more."

Political orientation

Participants were asked "How do you identify your political orientation on *social* issues?" and "How do you identify your political orientation on *economic* issues?" For each question, participants selected one option from the following: Very Conservative (1), Somewhat

Conservative (2), Slightly Conservative (3), Moderate/Middle of the Road (4), Slightly Liberal (5), Somewhat Liberal (6), Very Liberal (7).

Party affiliation

Participants were asked "How do you identify your political party affiliation?" and selected on of the following options: Strong Republican (1), Republican (2), Lean Republican (3), Neither Republican or Democrat (4), Lean Democrat (5), Democrat (6), Strong Democrat (7).

Moral conviction descriptives

In Study 1a, the measure of moral conviction had suitable internal reliability (Cronbach's $\alpha = 0.74$), and participants felt morally convicted about the issue on average ($M = 4.71$, $SD = 1.44$). The measure of moral conviction also had suitable internal reliability in Study 1b as well (Cronbach's $\alpha = 0.82$), and participants felt moderately convicted about trigger warnings on average ($M = 4.11$, $SD = 1.44$).

Condition differences in evaluations

I conducted Welch's t-tests to assess whether participants in the blinded condition evaluated the presented studies more positively than those in the unblinded conditions (Delacre et al., 2017)[8]. As hypothesized, participants evaluated the study as being of significantly higher quality in the blinded condition ($M = 5.01$, $SD = 0.93$) compared to the unblinded condition ($M = 4.60$, $SD = 1.17$) in Study 1a, Welch's $t(440.32) = 4.20$, $p < .001$, Cohen's $d = 0.39$. However, this was not the case in Study 1b, as those in the blinded condition ($M = 5.07$, $SD = 0.96$) and unblinded condition ($M = 5.03$, $SD = 1.12$) did not significantly differ in quality evaluations, Welch's $t(446.01) = 0.39$, $p = .35$, Cohen's $d = 0.04$.

[8] Following my preregistration, I used one-tailed t-tests for these analyses to assess my directional hypotheses.

Political orientation and party affiliation exploratory analyses

Participants reported being somewhat liberal on social (M = 5.44, SD = 1.65) and economic (M = 5.08, SD = 1.76) issues and leaned Democrat (M = 5.08, SD = 1.54) in Study 1a. Similarly, Participants reported being somewhat liberal on social (M = 5.18, SD = 1.74) and economic (M = 4.78, SD = 1.81) issues and leaned Democrat (M = 4.91, SD = 1.57) in Study 1b. Given the suitable reliability of the political orientation measures in Study 1a (Cronbach's α = 0.88) and Study 1b (Cronbach's α = 0.88), these items were combined into a composite political orientation measures in the following analyses.

Study 1a

For the analysis using political orientation as the predictor, omnibus tests showed a significant main effect of condition $F(1, 462) = 17.47$, $p < .001$, $\eta_p^2 = .04$, but not of political orientation, $F(1, 462) = 2.05$, $p = .15$, $\eta_p^2 = .00$. However, this was qualified by a significant interaction, $F(1, 462) = 11.78$, $p < .001$, $\eta_p^2 = .02$. Simple effects analyses showed that more liberally-orientated participants had lower study quality evaluations in the unblinded condition, b = -0.15, SE = 0.04, $p < .001$, 95% CI [-0.23, -0.06], β = -0.22, but not in the blinded condition, b = 0.06, SE = 0.04, $p = .15$, 95% CI [-0.02, 0.14], β = 0.09.

The model using party affiliation was yielded a weaker, but highly similar, interaction effect, $F(1, 462) = 6.11$, $p = .014$, $\eta_p^2 = .01$. Simple effects analyses showed that more Democratic participants had lower study quality evaluations in the unblinded condition, b = -0.09, SE = 0.05, $p = .047$, 95% CI [-0.18, -0.00], β = -0.13, but Democratic participants had nonsignificantly higher ratings of study quality in the blinded condition, b = 0.07, SE = 0.04, p = .14, 95% CI [-0.02, 0.15], β = 0.09.

Study 1b

Unlike in Study 1a, the analysis using the political orientation measure did not find a significant main effect of condition $F(1, 459) = 0.22$, $p = .64$, $\eta_p^2 = .00$, political orientation, $F(1, 459) = 0.50$, $p = .48$, $\eta_p^2 = .00$, nor a significant interaction, $F(1, 459) = 2.26$, $p = .13$, $\eta_p^2 = .00$. Simple effects analyses showed that, as in Study 1a, being more liberally-orientated was significantly predictive of lower study quality evaluations in the unblinded condition, $b = -0.12$, $SE = 0.04$, $p = .003$, 95% CI [-0.19, -0.04], $\beta = -0.19$, but not in the blinded condition, $b = -0.03$, $SE = 0.04$, $p = .48$, 95% CI [-0.11, 0.05], $\beta = -0.05$. However, the differences between the blinded and unblinded estimates were not significant in this study. A similar result emerged for the analysis of party affiliation, though the interaction term was closer to statistical significance in this model, $F(1, 459) = 3.29$, $p = .07$, $\eta_p^2 = .01$. More Democratic participants had significantly lower study quality evaluations in the unblinded condition, $b = -0.13$, $SE = 0.04$, $p = .047$, 95% CI [-0.21, -0.04], $\beta = -0.19$, but not in the blinded condition, $b = -0.01$, $SE = 0.05$, $p = .74$, 95% CI [-0.10, 0.07], $\beta = -0.02$. Though these results were similar to the analysis of party affiliation in Study 1a, the difference between the blinded and unblinded estimates was not significant in Study 1b.

Credibility impressions

Study 1b prior efficacy models

The main effect of condition was not significant, $F(1, 459) = 0.76$, $p = .38$, $\eta_p^2 = .00$, and neither was the main effect of prior efficacy beliefs, $F(1, 459) = 0.91$, $p = .34$, $\eta_p^2 = .00$, nor the interaction, $F(1, 459) = 0.02$, $p = .89$, $\eta_p^2 = .00$. The results of the moderated mediation analysis are presented in Table S1.1.

Reported attitude change analyses

Reported change in support beliefs

Study 1a. The intercept for the regression model of Study 1a (b = 4.14, SE = 0.05, 95% CI [4.04, 4.23]) indicated that participants reported a slight increase in their support for capital punishment on average. There was not a main effect of condition, $F(1, 462) = 0.03$, $p = .85$, $\eta_p^2 = .00$, but there was a large main effect or prior support, $F(1, 462) = 12.01$, $p < .001$, $\eta_p^2 = .03$, such that prior support for capital punishment was predictive of greater reported attitude change, b = 0.09, SE = 0.02, 95% CI [0.04, 0.13], $\beta = 0.22$. This result went in the opposite direction of the analysis of actual attitude change, in which those with greater prior support actually changed their attitudes *less*. The interaction was not significant, $F(1, 462) = 0.00$, $p = .96$, $\eta_p^2 = .00$, indicating that the lack of effect of condition on reported change in support beliefs did not vary as a function of participants' initial support beliefs.

Nevertheless, as was the case in the analysis of actual change in support, the moderated mediation model showed that, while the amount of reported belief updating that occurred did not vary by condition or participants' prior support, the processes by which participants reported updating their support beliefs did significantly vary across these factors. The estimates of the moderated mediation model are presented in Table S1.2. The results were nearly identical to those with actual belief change as the outcome: positive direct effects of condition on reported support change negated the negative indirect effects of condition on support change through quality evaluations, but only for participants with average or below-average prior support beliefs—the participants for whom the presented results were politically-unfriendly.

Study 1b. For Study 1b, the intercept for the regression model (b = 3.49, SE = 0.07, 95% CI [3.36, 3.63]) showed that participants reported a small decrease in their support for trigger warnings on average. There were not significant main effects of condition, $F(1, 459) = 0.01$, $p = .94$, $\eta_p^2 = .00$, nor prior support, $F(1, 459) = 1.00$, $p = .32$, $\eta_p^2 = .00$, and there was also not a

significant interaction, $F(1, 459) = 2.44$, p .32, $\eta_p^2 = .01$. Moreover, the results of the moderated mediation analysis, presented in Table S1.3, indicated that there were no significant indirect, direct, or total effects of condition on reported change in support.

Reported change in efficacy beliefs

Study 1a. The intercept for the regression model of Study 1a ($b = 4.23$, $SE = 0.12$, 95% CI [3.99, 4.47]) indicated that participants reported not changing their beliefs about the efficacy of capital punishment on average. There was not a main effect of condition, $F(1, 462) = 0.51$, $p = .48$, $\eta_p^2 = .00$, nor prior support, $F(1, 462) = 0.45$, $p = .51$, $\eta_p^2 = .00$, nor a significant interaction, $F(1, 462) = 2.11$, $p = .15$, $\eta_p^2 = .00$. However, as was the case for the analysis of actual change in support beliefs for Study 1a, the blinding manipulation had different effects on reported change in efficacy beliefs depending on participants' level of prior efficacy beliefs. The results of the moderated mediation analysis, presented in Table S1.4, showed that there were countervailing indirect effects through study quality evaluations and direct effects of condition for participants with average and below-average prior efficacy beliefs. This was not the case for participants with above-average prior efficacy beliefs, for whom there were no indirect, direct, or total effects of condition on changes in reported efficacy beliefs.

Study 1b. For Study 1b, the intercept for the regression model ($b = 3.18$, $SE = 0.05$, 95% CI [3.02, 3.34]) showed that participants reported a decrease in their support for capital punishment on average. There were not significant main effects of condition, $F(1, 459) = 0.10$, $p = .76$, $\eta_p^2 = .00$, nor prior support, $F(1, 459) = 0.46$, $p = .50$, $\eta_p^2 = .00$, but there was a small, significant interaction, $F(1, 459) = 3.89$, $p = .049$, $\eta_p^2 = .01$. This interaction indicated that participants with above-average prior efficacy beliefs reported (nonsignificantly) more change in efficacy beliefs in the unblinded condition, and those with below-average efficacy beliefs had the

opposite pattern of results. However, the results of the moderated mediation analysis, presented in Table S1.5, indicated that there were no significant indirect, direct, or total effects of condition on reported change in efficacy beliefs for Study 1b, as was the case for reported change in support beliefs for this study. .

Table S1.1. *Moderated mediation estimates of condition predicting credibility impressions by prior efficacy beliefs for Study 1b.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Low (M – 1SD)	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.08	0.10	[-0.11, 0.27]	0.04	.38
		Condition ⇒ Study Quality	0.12	0.14	[-0.16, 0.40]	0.06	.39
		Study Quality ⇒ Credibility impressions	0.67	0.05	[0.59, 0.77]	0.75	< .001
	Direct	Condition ⇒ Credibility impressions	-0.13	0.08	[-0.30, 0.02]	-0.07	.10
	Total	Condition ⇒ Credibility impressions	-0.07	0.13	[-0.33, 0.19]	-0.03	.60
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.02	0.07	[-0.16, 0.12]	-0.01	.74
		Condition ⇒ Study Quality	-0.03	0.09	[-0.21, 0.16]	-0.01	.74
		Study Quality ⇒ Credibility impressions	0.75	0.03	[0.70, 0.81]	0.78	< .001
	Direct	Condition ⇒ Credibility impressions	-0.05	0.06	[-0.16, 0.07]	-0.02	.43
	Total	Condition ⇒ Credibility impressions	-0.08	0.09	[-0.26, 0.10]	-0.04	.38
High (M + 1SD)	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.15	0.12	[-0.39, 0.09]	-0.07	.21
		Condition ⇒ Study Quality	-0.19	0.15	[-0.48, 0.10]	-0.09	.21
		Study Quality ⇒ Credibility impressions	0.82	0.04	[0.76, 0.90]	0.81	< .001
	Direct	Condition ⇒ Credibility impressions	0.04	0.08	[-0.12, 0.21]	0.02	.63
	Total	Condition ⇒ Credibility impressions	-0.09	0.13	[-0.35, 0.16]	-0.05	.48

Note. Condition was dummy-coded (0 = Blinded). Confidence intervals computed with 1000 bootstrap replications.

Table S1.2. *Moderated mediation estimates of condition predicting reported change in support beliefs by prior support beliefs for Study 1a.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.13	0.05	[-0.23, -0.02]	-0.08	.02
		Condition ⇒ Study Quality	-0.93	0.14	[-1.20, -0.66]	-0.43	< .001
		Study Quality ⇒ Reported Change in Support	0.14	0.05	[0.03, 0.24]	0.19	.009
	Direct	Condition ⇒ Reported Change in Support	0.12	0.09	[-0.07, 0.30]	0.08	.20
	Total	Condition ⇒ Reported Change in Support	-0.01	0.10	[-0.20, 0.18]	0.00	.95
Average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.06	0.02	[-0.11, -0.02]	-0.04	.004
		Condition ⇒ Study Quality	-0.38	0.09	[-0.57, -0.20]	-0.18	< .001
		Study Quality ⇒ Reported Change in Support	0.16	0.04	[0.09, 0.25]	0.23	< .001
	Direct	Condition ⇒ Reported Change in Support	0.06	0.07	[-0.07, 0.20]	0.04	.38
	Total	Condition ⇒ Reported Change in Support	0.01	0.07	[-0.12, 0.15]	0.01	.85
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.03	0.03	[-0.02, 0.09]	0.02	.27
		Condition ⇒ Study Quality	0.16	0.13	[-0.08, 0.41]	0.08	.20
		Study Quality ⇒ Reported Change in Support	0.19	0.06	[0.07, 0.31]	0.27	.002
	Direct	Condition ⇒ Reported Change in Support	0.00	0.11	[-0.22, 0.23]	0.00	.99
	Total	Condition ⇒ Reported Change in Support	0.03	0.10	[-0.16, 0.22]	0.02	.74

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S1.3. *Moderated mediation estimates of condition predicting reported change in support beliefs by prior support beliefs for Study 1b.*

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.05	0.04	[-0.14, 0.03]	-0.02	.21
		Condition ⇒ Study Quality	0.19	0.13	[-0.07, 0.44]	0.09	.15
		Study Quality ⇒ Reported Change in Support	-0.28	0.08	[-0.44, -0.12]	-0.27	< .001
	Direct	Condition ⇒ Reported Change in Support	-0.09	0.16	[-0.40, 0.21]	-0.04	.58
	Total	Condition ⇒ Reported Change in Support	-0.15	0.14	[-0.42, 0.13]	-0.07	.29
Average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.01	0.03	[-0.04, 0.06]	0.01	.63
		Condition ⇒ Study Quality	-0.05	0.10	[-0.25, 0.15]	-0.02	.62
		Study Quality ⇒ Reported Change in Support	-0.24	0.05	[-0.34, -0.14]	-0.24	< .001
	Direct	Condition ⇒ Reported Change in Support	0.00	0.10	[-0.19, 0.19]	0.00	.97
	Total	Condition ⇒ Reported Change in Support	0.01	0.10	[-0.19, 0.20]	0.00	.94
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.06	0.04	[-0.02, 0.13]	0.03	.13
		Condition ⇒ Study Quality	-0.29	0.15	[-0.58, 0.02]	-0.14	.06
		Study Quality ⇒ Reported Change in Support	-0.20	0.07	[-0.34, -0.07]	-0.20	.003
	Direct	Condition ⇒ Reported Change in Support	0.09	0.15	[-0.18, 0.38]	0.04	.52
	Total	Condition ⇒ Reported Change in Support	0.16	0.14	[-0.11, 0.43]	0.08	.24

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S1.4. *Moderated mediation estimates of condition predicting reported change in efficacy beliefs by prior efficacy beliefs for Study 1a.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.26	0.07	[-0.40, -0.13]	-0.14	< .001
		Condition ⇒ Study Quality	-0.97	0.14	[-1.24, -0.70]	-0.45	< .001
		Study Quality ⇒ Reported Change in Support	0.27	0.06	[0.16, 0.38]	0.31	< .001
	Direct	Condition ⇒ Reported Change in Support	0.23	0.11	[0.00, 0.44]	0.12	.047
	Total	Condition ⇒ Reported Change in Support	-0.03	0.12	[-0.27, 0.21]	-0.02	.79
Average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.09	0.03	[-0.15, -0.03]	-0.05	.002
		Condition ⇒ Study Quality	-0.38	0.09	[-0.56, -0.20]	-0.18	< .001
		Study Quality ⇒ Reported Change in Support	0.24	0.05	[0.14, 0.33]	0.27	< .001
	Direct	Condition ⇒ Reported Change in Support	0.20	0.09	[0.03, 0.37]	0.11	.022
	Total	Condition ⇒ Reported Change in Support	0.09	0.09	[-0.08, 0.26]	0.05	.28
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.04	0.03	[-0.02, 0.10]	0.02	.17
		Condition ⇒ Study Quality	0.21	0.13	[-0.04, 0.46]	0.10	.099
		Study Quality ⇒ Reported Change in Support	0.20	0.08	[0.05, 0.36]	0.24	.008
	Direct	Condition ⇒ Reported Change in Support	0.18	0.14	[-0.09, 0.46]	0.10	.20
	Total	Condition ⇒ Reported Change in Support	0.22	0.12	[-0.02, 0.46]	0.12	.073

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S1.5. *Moderated mediation estimates of condition predicting reported change in efficacy beliefs by prior efficacy beliefs for Study 1b.*

Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	-0.03	0.04	[-0.12, 0.05]	-0.01	.42
		Condition ⇒ Study Quality	0.12	0.15	[-0.17, 0.41]	0.06	.39
		Study Quality ⇒ Reported Change in Support	-0.27	0.07	[-0.41, -0.13]	-0.23	< .001
	Direct	Condition ⇒ Reported Change in Support	-0.15	0.16	[-0.46, 0.16]	-0.06	.35
	Total	Condition ⇒ Reported Change in Support	-0.19	0.16	[-0.51, 0.13]	-0.08	.24
Average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.01	0.02	[-0.04, 0.06]	0.00	.75
		Condition ⇒ Study Quality	-0.03	0.09	[-0.22, 0.15]	-0.01	.74
		Study Quality ⇒ Reported Change in Support	-0.24	0.06	[-0.35, -0.13]	-0.21	< .001
	Direct	Condition ⇒ Reported Change in Support	0.03	0.11	[-0.19, 0.26]	0.01	.78
	Total	Condition ⇒ Reported Change in Support	0.04	0.11	[-0.19, 0.26]	0.01	.75
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Reported Change in Support	0.04	0.04	[-0.03, 0.11]	0.02	.27
		Condition ⇒ Study Quality	-0.19	0.15	[-0.48, 0.10]	-0.09	.20
		Study Quality ⇒ Reported Change in Support	-0.21	0.08	[-0.37, -0.06]	-0.18	.008
	Direct	Condition ⇒ Reported Change in Support	0.21	0.18	[-0.14, 0.56]	0.09	.23
	Total	Condition ⇒ Reported Change in Support	0.26	0.16	[-0.05, 0.58]	0.11	.11

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

APPENDIX B: Study 2 Supplement

Study quality evaluations by condition

As hypothesized, participants evaluated the study as being of significantly higher quality in the blinded condition (M = 5.21, SD = 0.93) compared to the unblinded condition (M = 4.72, SD = 1.15), Welch's $t(877.54) = 6.97$, $p < .001$, Cohen's $d = 0.46$.

Exploratory analysis of political orientation

This model predicted study quality evaluations from condition, political orientation, and their interaction. As in the confirmatory analyses, this two-way interaction was significant, $F(1, 908) = 22.38$, $p < .001$, $\eta_p^2 = .02$. Simple effects analyses showed that, while political orientation was not significantly related to quality evaluations in the blinded condition, b = 0.00, SE = 0.03, $p < .001$, 95% CI [-0.05, 0.05], β = 0.01, it was significantly related to quality judgments in the unblinded condition, b = -0.17, SE = 0.03, $p < .001$, 95% CI [-0.22, -0.12], β = -0.29. More liberal participants in the unblinded condition tended to rate the study quality as being lower, but political orientation was not predictive of quality evaluations in the blinded condition. Moreover, an analysis of the estimated marginal mans, presented in Table S2.1, indicated that more liberal participants—for whom the presented results were more politically unfriendly—had larger discrepancies between unblinded and blinded quality evaluations. The discrepancy between blinded and unblinded quality judgments of politically-similar participants provided further evidence that participants' political beliefs biased their unblinded quality evaluations.

Table S2.1. Estimated Marginal Means of Study Quality Evaluations in Study 2 by Condition and Political Orientation

	Condition	
Political Orientation Group	Blinded	Unblinded
Below-average (More Conservative)	5.20 [5.06, 5.34]	5.03 [4.90, 5.16]
Average	5.21 [5.11, 5.30]	**4.71 [4.62, 4.81]**
Above-average (More Liberal)	5.21 [5.08, 5.34]	**4.40 [4.26, 4.53]**

Note: Values in brackets are 95% confidence intervals. Bolded values indicate nonoverlapping confidence intervals from the Blinded condition.

Mediation of study quality judgments by positive and negative affect

The results of the moderated mediation analyses are presented in Tables S2.2 and S2.3.

Prior efficacy beliefs moderating the effect of condition on credibility impressions

As was the case in the model using prior support, there were significant main effects of condition, $F(1, 908) = 17.57$, $p < .001$, $\eta_p^2 = .02$, and prior support $F(1, 908) = 35.81$, $p < .001$, $\eta_p^2 = .04$, qualified by a significant interaction, $F(1, 908) = 18.36$, $p < .001$, $\eta_p^2 = .02$. Greater prior belief in the efficacy of capital punishment was predictive of more positive credibility impressions across conditions, but this association was significantly stronger in the unblinded condition, $b = 0.29$, $SE = 0.02$, $p < .001$, 95% CI [0.24, 0.34], $\beta = 0.51$, than in the blinded condition, $b = 0.14$, $SE = 0.02$, $p < .001$, 95% CI [0.10, 0.19], $\beta = 0.25$. The estimated marginal means, shown in Table S2.4, indicated that participants with average and below-average prior efficacy beliefs had significantly more positive credibility impressions in the blinded condition than in the unblinded condition, while participants with above-average prior support for capital punishment did not differ in their credibility impressions across conditions. The results of the moderated mediation analysis, presented in Table S2.5, showed that the differences in credibility impressions between conditions for participants with average and below-average prior efficacy beliefs could be explained by the influence of the biased quality evaluations made by those unblinded participants.

Table S2.4. Estimated Marginal Means of Credibility Impressions Evaluations in Study 2 by Condition and Prior Efficacy Beliefs

	Condition	
Prior Efficacy Group	Blinded	Unblinded
Below-average	4.62 [4.50, 4.75]	**4.07 [3.94, 4.20]**
Average	4.89 [4.80, 4.99]	**4.62 [4.53, 4.71]**
Above-average	5.17 [5.04, 5.29]	5.17 [5.04, 5.30]

Note: Values in brackets are 95% confidence intervals. Bolded values indicate nonoverlapping confidence intervals from the Blinded condition.

Reported changes in support and efficacy beliefs

Reported change in support

In the linear regression model predicting reported change in support for capital punishment, the intercept (b = 4.17, SE = 0.04, 95% CI [4.09, 4.25]) indicated that participants reported a small increase in their support on average. There was not a significant main effect of condition, $F(1, 908) = 1.04$, $p = .31$, $\eta_p^2 = .00$, but there was a significant main effect of prior support, $F(1, 908) = 37.66$, $p < .001$, $\eta_p^2 = .04$, such that participants who supported capital punishment more generally reported changing their support beliefs more, b = 0.12, SE = 0.02, 95% CI [0.08, 0.16], β = 0.27. This ran counter to the analysis of actual change in support, as participants with higher prior support actually changed their beliefs less than other participants, likely because the presented results aligned with what they already believed. There was not a significant interaction between condition and prior support, $F(1, 908) = 0.12$, $p = .73$, $\eta_p^2 = .00$, indicating that the lack of influence of condition on reported change in support beliefs was consistent across levels of prior support.

Reported change in efficacy beliefs

The intercept of this model (b = 4.42, SE = 0.05, 95% CI [4.33, 4.51]) showed that participants reported an increase in their efficacy beliefs on average. There was not a significant main effect of condition, $F(1, 908) = 0.03$, $p = .85$, $\eta_p^2 = .00$, nor a main effect of prior efficacy

beliefs, $F(1, 908) = 3.71$, $p = .054$, $\eta_p^2 = .00$. However, there was a small, significant interaction between condition and prior efficacy beliefs, $F(1, 908) = 6.80$, $p = .009$, $\eta_p^2 = .01$, indicating that the influence of condition on reported change in efficacy beliefs differed across levels of prior support. Specifically, participants with below-average prior efficacy beliefs reported changing their beliefs slightly (but nonsignificantly) more in the blinded condition ($M = 4.34$, 95% CI [4.21, 4.46]) than in the unblinded condition ($M = 4.18$, 95% CI [4.05, 4.31]), whereas those with above-average prior efficacy beliefs reported changing their beliefs slightly (but nonsignificantly) *less* in the blinded condition ($M = 4.51$, 95% CI [4.38, 4.64]) than in the unblinded condition ($M = 4.69$, 95% CI [4.56, 4.81]).

Table S2.2. Moderated mediation estimates of condition predicting study quality by prior support for Study 2.

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.03	0.02	[-0.06, 0.00]	-0.01	.093
		Condition ⇒ Negative Affect ⇒ Study Quality	-0.01	0.02	[-0.04, 0.02]	-0.01	.43
	Component	Condition ⇒ Positive Affect	-0.15	0.08	[-0.30, 0.00]	-0.09	.051
		Positive Affect ⇒ Study Quality	0.19	0.06	[0.08, 0.32]	0.15	.002
		Condition ⇒ Negative Affect	0.04	0.05	[-0.05, 0.13]	0.04	.389
		Negative Affect ⇒ Study Quality	-0.32	0.11	[-0.53, -0.11]	-0.15	.003
	Direct	Condition ⇒ Study Quality	-1.05	0.09	[-1.23, -0.87]	-0.49	< .001
	Total	Condition ⇒ Study Quality	-1.09	0.09	[-1.27, -0.91]	-0.51	< .001
Average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.03	0.02	[-0.07, -0.00]	-0.02	.033
		Condition ⇒ Negative Affect ⇒ Study Quality	0.00	0.01	[-0.02, 0.01]	0.00	.72
	Component	Condition ⇒ Positive Affect	-0.13	0.06	[-0.24, -0.02]	-0.07	.021
		Positive Affect ⇒ Study Quality	0.27	0.04	[0.19, 0.35]	0.21	< .001
		Condition ⇒ Negative Affect	0.01	0.03	[-0.05, 0.08]	0.01	.70
		Negative Affect ⇒ Study Quality	-0.22	0.06	[-0.34, -0.10]	-0.11	< .001
	Direct	Condition ⇒ Study Quality	-0.47	0.06	[-0.59, -0.34]	-0.22	< .001
	Total	Condition ⇒ Study Quality	-0.51	0.06	[-0.64, -0.38]	-0.24	< .001
Above-average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.04	0.03	[-0.10, 0.02]	-0.02	.22
		Condition ⇒ Negative Affect ⇒ Study Quality	0.00	0.01	[-0.01, 0.02]	0.00	.80
	Component	Condition ⇒ Positive Affect	-0.11	0.08	[-0.27, 0.06]	-0.06	.20
		Positive Affect ⇒ Study Quality	0.34	0.04	[0.26, 0.43]	0.27	< .001
		Condition ⇒ Negative Affect	-0.01	0.05	[-0.11, 0.08]	-0.01	.78
		Negative Affect ⇒ Study Quality	-0.13	0.08	[-0.30, 0.03]	-0.06	.12
	Direct	Condition ⇒ Study Quality	0.12	0.08	[-0.05, 0.28]	0.05	.17
	Total	Condition ⇒ Study Quality	0.08	0.09	[-0.10, 0.26]	0.04	.39

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table 2.3. Moderated mediation estimates of condition predicting study quality by prior efficacy beliefs for Study 2.

Prior Support Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.03	0.02	[-0.06, 0.01]	-0.01	.11
		Condition ⇒ Negative Affect ⇒ Study Quality	0.00	0.02	[-0.04, 0.03]	0.00	.91
	Component	Condition ⇒ Positive Affect	-0.16	0.08	[-0.31, -0.01]	-0.09	.036
		Positive Affect ⇒ Study Quality	0.17	0.06	[0.05, 0.29]	0.13	.008
		Condition ⇒ Negative Affect	0.01	0.05	[-0.08, 0.10]	0.01	.91
		Negative Affect ⇒ Study Quality	-0.39	0.10	[-0.59, -0.20]	-0.19	< .001
	Direct	Condition ⇒ Study Quality	-1.08	0.09	[-1.26, -0.90]	-0.50	< .001
	Total	Condition ⇒ Study Quality	-1.11	0.09	[-1.29, -0.94]	-0.52	< .001
Average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.03	0.01	[-0.05, -0.00]	-0.01	.041
		Condition ⇒ Negative Affect ⇒ Study Quality	0.00	0.01	[-0.02, 0.02]	0.00	.76
	Component	Condition ⇒ Positive Affect	-0.12	0.05	[-0.22, -0.01]	-0.07	.025
		Positive Affect ⇒ Study Quality	0.23	0.04	[0.15, 0.32]	0.19	< .001
		Condition ⇒ Negative Affect	0.01	0.04	[-0.06, 0.08]	0.01	.75
		Negative Affect ⇒ Study Quality	-0.27	0.06	[-0.39, -0.15]	-0.13	< .001
	Direct	Condition ⇒ Study Quality	-0.47	0.06	[-0.59, -0.34]	-0.22	< .001
	Total	Condition ⇒ Study Quality	-0.49	0.06	[-0.62, -0.37]	-0.23	< .001
Above-average	Indirect	Condition ⇒ Positive Affect ⇒ Study Quality	-0.02	0.03	[-0.07, 0.03]	-0.01	.39
		Condition ⇒ Negative Affect ⇒ Study Quality	0.00	0.01	[-0.02, 0.01]	0.00	.77
	Component	Condition ⇒ Positive Affect	-0.07	0.08	[-0.24, 0.09]	-0.04	.37
		Positive Affect ⇒ Study Quality	0.30	0.04	[0.21, 0.39]	0.24	< .001
		Condition ⇒ Negative Affect	0.02	0.05	[-0.08, 0.12]	0.02	.74
		Negative Affect ⇒ Study Quality	-0.15	0.07	[-0.28, -0.02]	-0.07	.031
	Direct	Condition ⇒ Study Quality	0.15	0.08	[-0.01, 0.31]	0.07	.063
	Total	Condition ⇒ Study Quality	0.12	0.09	[-0.05, 0.30]	0.06	.17

Note. Condition was dummy-coded (0 = Blinded). Below-average prior beliefs = M -1SD, above-average prior beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S2.5. Moderated mediation estimates of condition predicting credibility impressions by prior efficacy beliefs for Study 2.

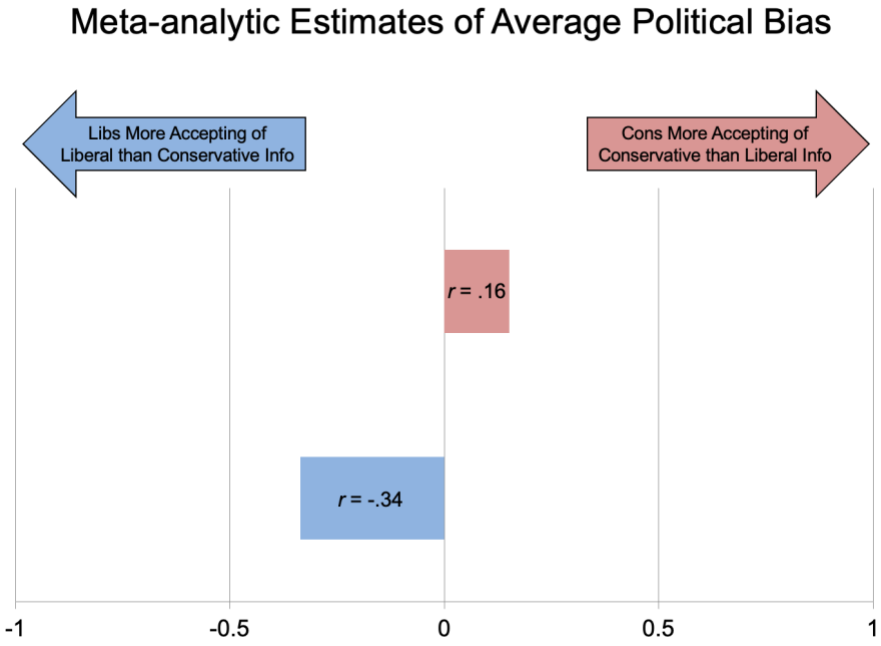
Prior Efficacy Group	Type	Effect	Estimate	SE	95% CI	β	p
Below-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.81	0.08	[-0.96, -0.66]	-0.38	< .001
		Condition ⇒ Study Quality	-1.11	0.09	[-1.30, -0.93]	-0.52	< .001
		Study Quality ⇒ Credibility impressions	0.73	0.03	[0.67, 0.79]	0.73	< .001
	Direct	Condition ⇒ Credibility impressions	0.26	0.07	[0.12, 0.40]	0.12	< .001
	Total	Condition ⇒ Credibility impressions	-0.55	0.09	[-0.74, -0.37]	-0.26	< .001
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.36	0.05	[-0.46, -0.27]	-0.17	< .001
		Condition ⇒ Study Quality	-0.49	0.06	[-0.62, -0.37]	-0.23	< .001
		Study Quality ⇒ Credibility impressions	0.73	0.03	[0.68, 0.79]	0.73	< .001
	Direct	Condition ⇒ Credibility impressions	0.09	0.05	[0.00, 0.18]	0.04	.064
	Total	Condition ⇒ Credibility impressions	-0.27	0.07	[-0.40, -0.15]	-0.13	< .001
Above-average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.09	0.07	[-0.04, 0.22]	0.04	.17
		Condition ⇒ Study Quality	0.12	0.09	[-0.05, 0.30]	0.06	.16
		Study Quality ⇒ Credibility impressions	0.73	0.04	[0.66, 0.81]	0.73	< .001
	Direct	Condition ⇒ Credibility impressions	-0.08	0.06	[-0.21, 0.04]	-0.04	.19
	Total	Condition ⇒ Credibility impressions	0.01	0.09	[-0.17, 0.19]	0.00	.94

Note. Condition was dummy-coded (0 = Blinded). Below-average prior efficacy beliefs = M -1SD, above-average prior efficacy beliefs = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

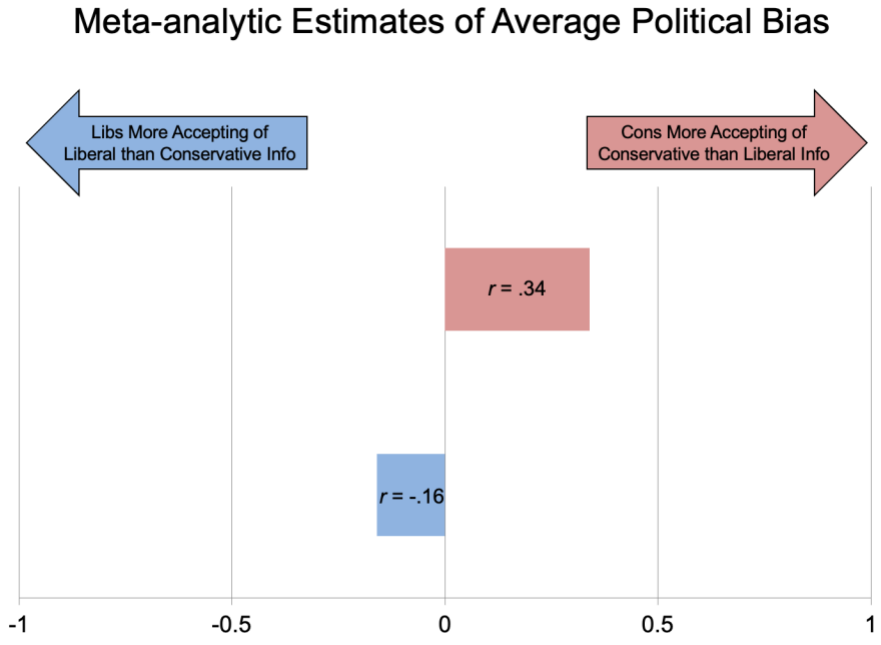
APPENDIX C: Study 3 Supplement

Figures accompanying the experimental manipulations

Conservative-friendly materials



Liberal-friendly materials



Influence of prior beliefs on study quality evaluations

Omnibus tests showed that the overall model was significant, $F(7, 1754) = 39.65$, $p < .001$, $\eta_p^2 = .14$, as were the main effects of condition, $F(1, 1754) = 19.43$, $p < .001$, $\eta_p^2 = .01$, prior beliefs, $F(1, 1754) = 18.14$, $p < .001$, $\eta_p^2 = .01$, but not of materials, $F(1, 1754) = 3.68$, $p = .055$, $\eta_p^2 = .00$. Yet the key three-way interaction term was significant, $F(1, 1754) = 95.28$, $p < .001$, $\eta_p^2 = .05$, as illustrated in Figure 3.3 in the main text. Prior beliefs positively predicted study quality evaluations in both the blinded conditions of the conservative-friendly materials, $b = 0.12$, $SE = 0.03$, $p < .001$, 95% CI [0.07, 0.18], $\beta = 0.19$, and the liberal-friendly materials, $b = 0.13$, $SE = 0.03$, $p < .001$, 95% CI [0.07, 0.18], $\beta = 0.20$. As was the case for partisan feelings, we observed an unexpected association between prior beliefs and study quality in the blinded conditions, such that participants with more liberal-friendly prior beliefs tended to evaluate the study more positively. However, the influence of prior beliefs on quality evaluations was much stronger in the unblinded conditions for both the conservative-friendly materials, $b = -0.22$, $SE = 0.03$, $p < .001$, 95% CI [-0.28, -0.16], $\beta = -0.34$, and the liberal-friendly materials, $b = 0.34$, $SE = 0.03$, $p < .001$, 95% CI [0.29, 0.39], $\beta = 0.53$. Unblinded participants with more liberal-leaning priors evaluated the study much more positively than those with more conservative-leaning priors in the liberal-friendly condition, yet unblinded participants with more liberal-leaning priors evaluated the study much more *negatively* than participants with more conservative-leaning priors in the conservative-friendly condition. Mirroring the analysis of partisan feelings, knowing the results of the study amplified the association between prior beliefs and study quality in the liberal-friendly condition, and it flipped the association between prior beliefs and study quality in the conservative-friendly condition. The difference between the unblinded and blinded estimates within the same materials condition is the magnitude of directional bias that partisan

feelings exerted on study quality evaluations. The significant differences (as indicated by nonoverlapping 95% CIs) between unblinded and blinded estimates within the same materials condition signifies that prior beliefs tended to bias unblinded participants quality evaluations.

Moreover, analyses of the estimated marginal means, depicted in Figure 3.4 in the main text, showed similar results as the analyses using partisan feelings. In the conservative-friendly materials, participants with conservative-leaning feelings in the unblinded condition provided slightly more positive study quality evaluations than their conservative-leaning counterparts in the blinded condition. However, unlike in the analysis of partisan feelings, this difference was not significant. Consistent with the prior analysis, those with liberal-leaning feelings provided significantly lower evaluations in the unblinded condition relative to the blinded condition. When assessing the liberal-friendly materials, those with conservative-leaning feelings provided significantly more positive study quality evaluations in the blinded condition than in the unblinded condition, and liberal-leaning participants provided slightly (but nonsignificantly) more positive quality evaluations in the unblinded condition. Participants with average partisan feelings were generally more consistent in their study quality evaluations across conditions, although they provided slightly more positive quality evaluations in the blinded conditions.

Exploratory analyses

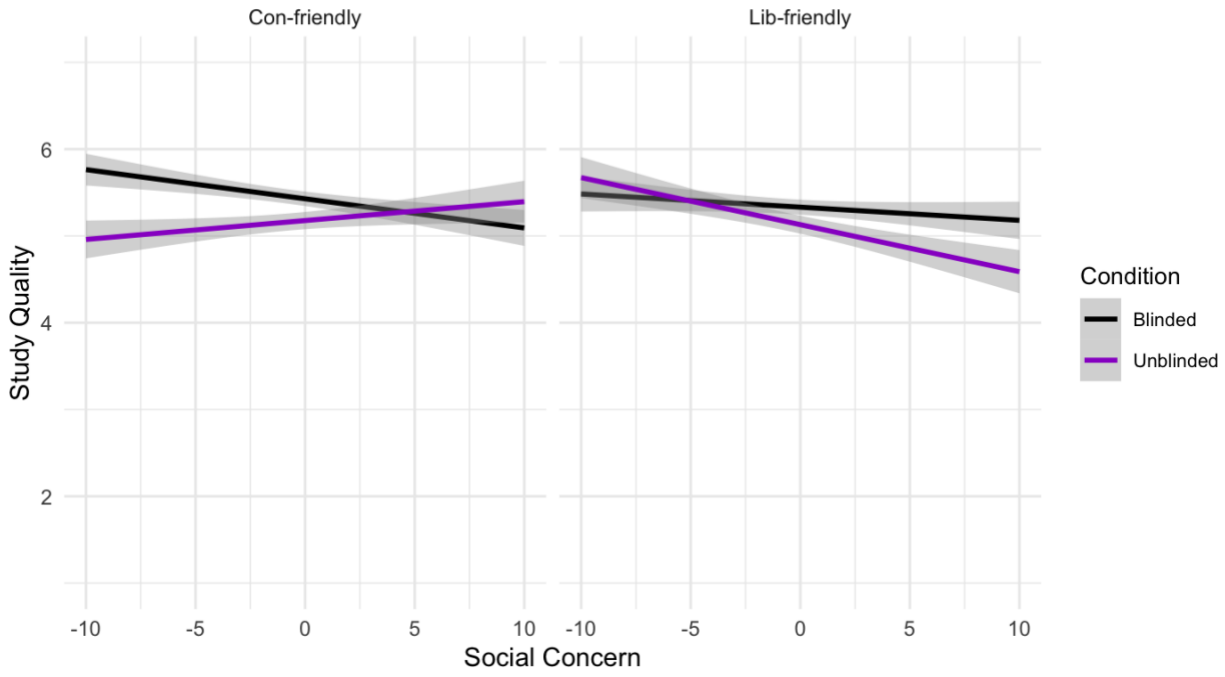
Influence of social concern

In line with the confirmatory analyses in Study 3, I constructed a regression model with the condition variable, (dummy-coded, 0 = blinded/conservative-friendly), the measure of social concern, and their interaction (results are substantively identical when including condition and materials as separate factors). Low scores on the social concern variable indicate perceptions of inhabiting a more liberal-friendly environment, and higher scores indicate perceptions of

inhabiting a more conservative-friendly environment. This model was constructed to examine how participants' sense of their social environment may have influenced their quality evaluations, which provides a more conservative test of the politically motivated reasoning model of partisan bias (Kahan, 2016).

Omnibus tests indicated that the overall model was significant, $F(7, 1754) = 10.82$, $p < .001$, $\eta_p^2 = .04$, as was the two main effects of condition and social concern ($ps \leq .001$). These main effects were further qualified by a significant interaction, $F(3, 1754) = 10.11$, $p < .001$, $\eta_p^2 = .02$. A simple effects analysis, depicted in Figure S3.1, showed that the magnitude of influence that participants' social concern had on their quality evaluations was significantly higher in the unblinded (compared to the blinded) conditions.

Figure S3.1. Average Study Quality Evaluations by Social Concern, Materials, and Condition in Study 3
Error bands represent standard errors



Furthermore, participants with low ($M - 1SD$) social concern scores (i.e., more liberal participants) rated the study as being lower quality in the unblinded/conservative-friendly condition compared to the other conditions, and participants with high ($M + 1SD$) social concern

scores (i.e., more conservative participants) rated the study as being lower quality in the unblinded/liberal-friendly condition compared to the other conditions. These estimated marginal means are shown in Table S3.1.

Table S3.1. *Estimated marginal means of study quality evaluations by condition and social concern*

Condition	Social Concern Group	*M*	SE	95% CI
Blinded Con-friendly	Liberal-leaning	5.60	0.05	[5.47, 5.73]
	Average	5.45	0.03	[5.35, 5.54]
	Conservative-leaning	5.29	0.03	[5.15, 5.43]
Blinded Lib-friendly	Liberal-leaning	5.41	0.05	[5.28, 5.54]
	Average	5.34	0.03	[5.25, 5.43]
	Conservative-leaning	5.27	0.03	[5.13 ,5.40]
Unblinded Con-friendly	Liberal-leaning	5.06	0.05	[4.93, 5.20]
	Average	5.17	0.03	[5.07, -0.15]
	Conservative-leaning	5.27	0.04	[5.14, 5.40]
Unblinded Lib-friendly	Liberal-leaning	5.41	0.05	[5.28, 5.54]
	Average	5.16	0.03	[5.07, 5.25]
	Conservative-leaning	4.91	0.03	[4.78, 5.03]

Note: Social concern: Liberal-leaning = M - $1SD$, Conservative-leaning = M + $1SD$.

In sum, the perceived political leanings of one's social environment influenced participants' evaluation of an empirical study, and those perceptions biased participants toward making more negative quality evaluations when they made unblinded evaluations of politically-unfriendly information. It is important to note that, while this effect was much smaller (in terms of variance explained) than the effects using partisan feelings and prior beliefs as the predictors, it was entirely consistent with those results. These findings further underscore the biasing influence that directional motivations had on study quality evaluations in the unblinded conditions of this experiment.

Influence of partisanship and strength of partisan identification

To assess the influence that partisanship and partisans' strength of party identification had on study quality evaluations, I constructed a linear regression model using the condition variable, (dummy-coded, 0 = blinded/conservative-friendly), a variable of whether participants were Democrat or Republican (dummy-coded, 0 = Democrat), the measure of strength of partisan identification (Bankert et al., 2017), the three two-way interaction terms between these variables, and the three-way interaction term (results are substantively identical when including condition and materials as separate factors).

The overall model was statistically significant, $F(15, 1395) = 16.22$, $p < .001$, $\eta_p^2 = .15$, as was the key three-way interaction term, $F(3, 1395) = 4.19$, $p = .006$, $\eta_p^2 = .01$. This interaction is depicted in Table S3.2. At low levels of partisan identification, Democrats did not demonstrate a bias in quality judgments, as the estimated marginal means did not significantly differ for weakly-identified Democrats between conditions. This was not true for Republicans, as weakly-identified Republicans still rated the study as being of lower quality in the unblinded/liberal-friendly condition (compared to the three other conditions). However, both Democrats and Republicans demonstrated partisan biases in quality evaluations at average and high levels of identification. Democrats in the unblinded/conservative-friendly conditions had significantly lower quality evaluations than similarly-identified Democrats in the other conditions, and Republicans in the unblinded/liberal-friendly condition rated study quality as being significantly lower than similarly-identified Republicans in the other conditions. Additionally, Republicans with average or high levels of party identification who were in the unblinded/conservative-friendly condition rated study quality as being significantly higher than similarly-identified partisans in the blinded conditions. Overall, while both Democrats and Republicans tended to become increasingly biased toward counter-attitudinal information at higher levels of partisan

identification, Republicans were also more biased toward counter-attitudinal information at low levels of party identification and more biased toward pro-attitudinal information at average and high levels of party identification.

Table 3.2. *Estimated marginal means of study quality evaluations by condition, party, and strength of partisan identification*

Condition	Party	Strength of Identification	*M*	SE	95% CI
Blinded Con-friendly	Democrat	Below-average	5.52	0.09	[5.34, 5.70]
		Average	5.63	0.07	[5.49, 5.77]
		Above-average	5.74	0.11	[5.53, 5.95]
	Republican	Below-average	5.28	0.11	[5.07, 5.50]
		Average	5.26	0.07	[5.11, 5.40]
		Above-average	5.23	0.10	[5.03, 5.43]
Blinded Lib-friendly	Democrat	Below-average	5.37	0.10	[5.19, 5.56]
		Average	5.53	0.07	[5.40, 5.65]
		Above-average	5.68	0.09	[5.50, 5.86]
	Republican	Below-average	5.09	0.10	[4.90, 5.28]
		Average	5.19	0.07	[5.05, 5.33]
		Above-average	5.29	0.10	[5.03, 5.43]
Unblinded Con-friendly	Democrat	Below-average	5.20	0.10	[5.00, 5.40]
		Average	5.00	0.07	[4.86, 5.14]
		Above-average	4.80	0.11	[4.59, 5.01]
	Republican	Below-average	5.44	0.11	[5.23, 5.66]
		Average	5.58	0.08	[5.43, 5.73]
		Above-average	5.71	0.10	[5.51, 5.91]
Unblinded Lib-friendly	Democrat	Below-average	5.58	0.09	[5.39, 5.76]
		Average	5.72	0.07	[5.59, 5.85]
		Above-average	5.86	0.10	[5.67, 6.05]
	Republican	Below-average	4.44	0.11	[4.22, 4.65]
		Average	4.51	0.07	[4.37, 4.66]
		Above-average	4.59	0.10	[4.40, 4.78]

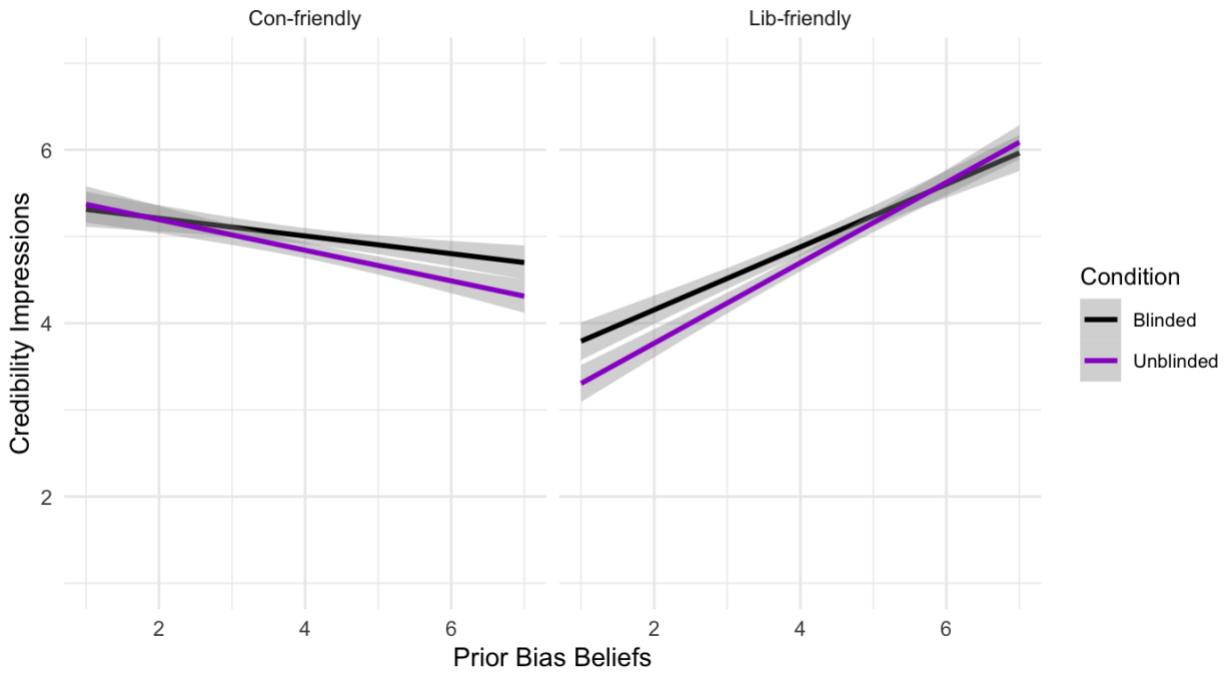
Note: Strength of identification: Below-average = *M* - *1SD*, Above-average = *M* + 1*SD*.

Influence of prior beliefs on credibility impressions

Omnibus tests showed that the overall model was significant, $F(7, 1754) = 63.94$, $p < .001$, $\eta_p^2 = .20$, as was the key three-way interaction term, $F(1, 1754) = 8.14$, $p = .004$, $\eta_p^2 = .00$, illustrated in Figure S3.2. A simple effects analysis showed that prior beliefs were more

predictive of credibility impressions in the unblinded conditions. In the conservative-friendly materials, having more liberal-leaning priors predicted negative credibility impressions more strongly in the unblinded condition, b = -0.18, SE = 0.03, p < .001, 95% CI [-0.24, -0.11], β = -0.24, than in the blinded condition, b = -0.10, SE = 0.03, p = .001, 95% CI [-0.16, -0.04], β = -0.14; in the liberal-friendly materials, having more liberal-leaning priors predicted positive credibility impressions more strongly in the unblinded condition, b = 0.46, SE = 0.03, p < .001, 95% CI [0.41, 0.52], β = 0.52, than in the blinded condition, b = 0.36, SE = 0.03, p < .001, 95% CI [0.30, 0.42], β = 0.42.

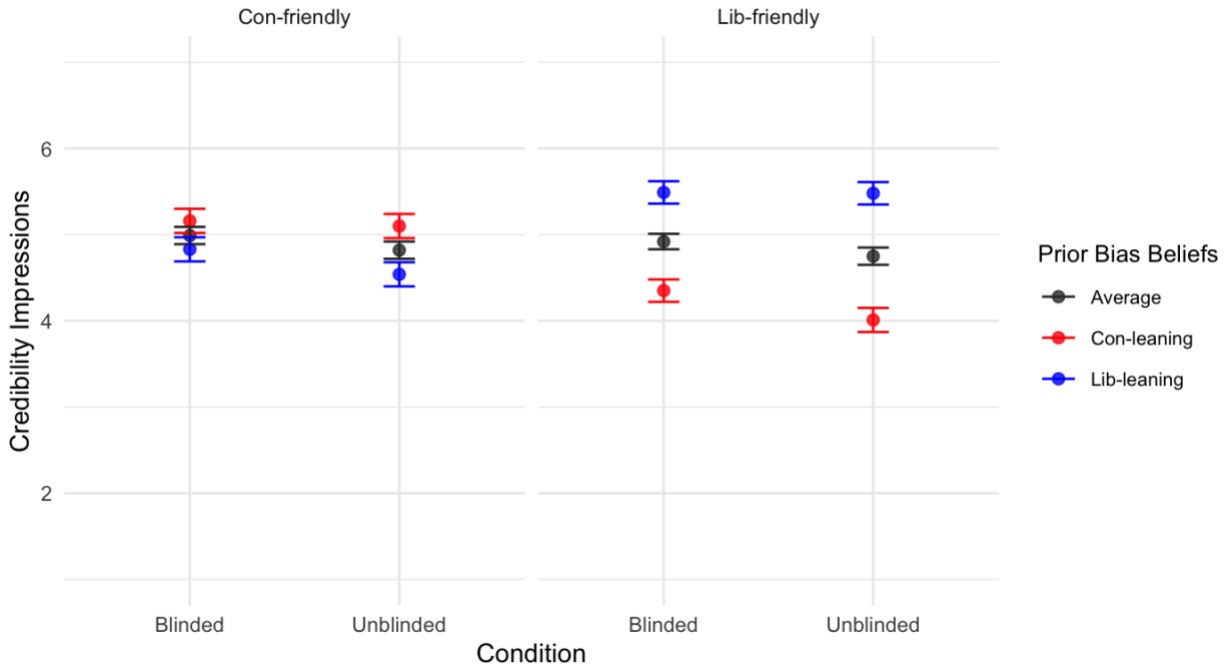
Figure S3.2. Average Credibility Impressions by Prior Bias Beliefs, Materials, and Condition in Study 3
Error bands represent standard errors



Moreover, Figure S3.3 depicts the estimated marginal means for this analysis. As was the case in the analysis of partisan feelings, blinding participants' quality evaluations made them form slightly more positive credibility impressions of politically-unfriendly information, yet all

participants tended to have more positive impressions of a study that produced politically-friendly results.

Figure S3.3. Average Credibility Impressions by Condition, Materials, and Prior Bias Beliefs in
Error bars represent 95% CIs



Tables S3.3 and S3.4 provide the results of moderated mediation analyses, separated by materials, that included condition (dummy-coded, 0 = blinded) as the predictor, study quality as the mediator, partisan feelings (mean-centered) as the moderator, and credibility impressions as the outcome variable. The pattern of results for this model were substantively identical to the model using partisan feelings as the moderator. The blinding manipulation reduced the influence that participants' quality evaluations had on their credibility impressions of politically-unfriendly information, which generally led to more positive credibility impressions in the blinded conditions.

Updating partisan feelings and beliefs about bias

Tables S3.5 – S3.8 present the results of the moderated mediation analyses assessing how blinding influenced belief updating across levels of participants' partisan feelings and prior bias beliefs.

Table S3.3. Moderated mediation estimates of condition predicting credibility impressions by prior beliefs in the lib-friendly materials for Study 3.

Prior Bias Beliefs Group	Type	Effect	Estimate	SE	95% CI	*β*	*p*
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.47	0.09	[-0.65, -0.29]	-0.18	< .001
		Condition ⇒ Study Quality	-0.55	0.11	[-0.76, -0.34]	-0.26	< .001
		Study Quality ⇒ Credibility impressions	0.86	0.04	[0.78, 0.93]	0.71	< .001
	Direct	Condition ⇒ Credibility impressions	0.14	0.08	[-0.01, 0.28]	0.05	.066
	Total	Condition ⇒ Credibility impressions	-0.33	0.10	[-0.53, -0.14]	-0.13	< .001
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.17	0.05	[-0.27, -0.07]	-0.07	.001
		Condition ⇒ Study Quality	-0.20	0.06	[-0.33, -0.08]	-0.10	.001
		Study Quality ⇒ Credibility impressions	0.83	0.03	[0.78, 0.89]	0.70	< .001
	Direct	Condition ⇒ Credibility impressions	0.01	0.05	[-0.08, 0.10]	0.00	.85
	Total	Condition ⇒ Credibility impressions	-0.17	0.07	[-0.31, -0.03]	-0.07	.016
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.11	0.07	[-0.02, 0.25]	0.04	.096
		Condition ⇒ Study Quality	0.14	0.08	[-0.02, 0.30]	0.07	.094
		Study Quality ⇒ Credibility impressions	0.81	0.03	[0.75, 0.88]	0.68	< .001
	Direct	Condition ⇒ Credibility impressions	-0.12	0.06	[-0.23, -0.01]	-0.05	.027
	Total	Condition ⇒ Credibility impressions	-0.01	0.10	[-0.20, 0.19]	0.00	.95

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S3.4. Moderated mediation estimates of condition predicting credibility impressions by prior bias beliefs in the con-friendly materials for Study 3.

Prior Bias Beliefs Group	Type	Effect	Estimate	SE	95% CI	β	p
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	0.18	0.07	[0.04, 0.32]	0.09	.009
		Condition ⇒ Study Quality	0.25	0.10	[0.06, 0.43]	0.13	.009
		Study Quality ⇒ Credibility impressions	0.73	0.03	[0.67, 0.79]	0.72	< .001
	Direct	Condition ⇒ Credibility impressions	-0.23	0.06	[-0.35, -0.10]	-0.12	< .001
	Total	Condition ⇒ Credibility impressions	-0.06	0.09	[-0.24, 0.13]	-0.03	.54
Average	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.22	0.05	[-0.32, -0.12]	-0.11	< .001
		Condition ⇒ Study Quality	-0.28	0.06	[-0.41, -0.16]	-0.15	< .001
		Study Quality ⇒ Credibility impressions	0.77	0.02	[0.72, 0.81]	0.74	< .001
	Direct	Condition ⇒ Credibility impressions	0.06	0.05	[-0.03, 0.15]	0.03	.17
	Total	Condition ⇒ Credibility impressions	-0.17	0.07	[-0.30, -0.04]	-0.09	.009
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Credibility impressions	-0.65	0.08	[-0.80, -0.50]	-0.32	< .001
		Condition ⇒ Study Quality	-0.82	0.09	[-0.98, -0.64]	-0.43	< .001
		Study Quality ⇒ Credibility impressions	0.80	0.03	[0.73, 0.87]	0.76	< .001
	Direct	Condition ⇒ Credibility impressions	0.36	0.07	[0.21, 0.51]	0.18	< .001
	Total	Condition ⇒ Credibility impressions	-0.29	0.09	[-0.47, -0.11]	-0.15	.002

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S3.5. *Moderated mediation estimates of condition predicting change in partisan feelings by initial partisan feelings in the conservative-friendly materials of Study 3.*

Partisan Feelings Group	Type	Effect	Estimate	SE	95% CI	β	p
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	-0.04	0.02	[-0.08, 0.01]	-0.01	.084
		Condition ⇒ Study Quality	0.23	0.10	[0.04, 0.42]	0.12	.018
		Study Quality ⇒ Change in Feelings	-0.16	0.06	[-0.28, -0.05]	-0.12	.005
	Direct	Condition ⇒ Change in Feelings	0.12	0.13	[-0.15, 0.38]	0.04	.39
	Total	Condition ⇒ Change in Feelings	0.08	0.13	[-0.17, 0.33]	0.03	.53
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	0.05	0.02	[0.01, 0.09]	0.02	.008
		Condition ⇒ Study Quality	-0.29	0.06	[-0.41, -0.17]	-0.15	< .001
		Study Quality ⇒ Change in Feelings	-0.18	0.05	[-0.29, -0.08]	-0.13	< .001
	Direct	Condition ⇒ Change in Feelings	0.02	0.10	[-0.17, 0.20]	0.01	.87
	Total	Condition ⇒ Change in Feelings	0.08	0.09	[-0.10, 0.25]	0.03	.39
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	0.16	0.06	[0.05, 0.28]	0.06	.006
		Condition ⇒ Study Quality	-0.81	0.08	[-0.97, -0.64]	-0.42	< .001
		Study Quality ⇒ Change in Feelings	-0.20	0.07	[-0.33, -0.07]	-0.15	.003
	Direct	Condition ⇒ Change in Feelings	-0.08	0.14	[-0.35, 0.18]	-0.03	.54
	Total	Condition ⇒ Change in Feelings	0.08	0.13	[-0.17, 0.32]	0.03	.55

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S3.6. *Moderated mediation estimates of condition predicting change in partisan feelings by initial partisan feelings in the liberal-friendly materials of Study 3.*

Partisan Feelings Group	Type	Effect	Estimate	SE	95% CI	β	p
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	-0.11	0.04	[-0.20, -0.03]	-0.05	.007
		Condition ⇒ Study Quality	-0.60	0.10	[-0.80, -0.39]	-0.28	< .001
		Study Quality ⇒ Change in Feelings	0.19	0.06	[0.07, 0.32]	0.17	.003
	Direct	Condition ⇒ Change in Feelings	-0.06	0.13	[-0.32, 0.19]	-0.03	.64
	Total	Condition ⇒ Change in Feelings	-0.19	0.11	[-0.40, 0.03]	-0.08	.091
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	-0.02	0.01	[-0.05, -0.00]	-0.01	.044
		Condition ⇒ Study Quality	-0.19	0.06	[-0.32, -0.06]	-0.09	.004
		Study Quality ⇒ Change in Feelings	0.13	0.05	[0.04, 0.22]	0.12	.003
	Direct	Condition ⇒ Change in Feelings	-0.03	0.08	[-0.18, 0.13]	-0.01	.71
	Total	Condition ⇒ Change in Feelings	-0.08	0.08	[-0.23, 0.08]	-0.03	.32
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Feelings	0.02	0.01	[-0.01, 0.04]	0.01	.21
		Condition ⇒ Study Quality	0.22	0.08	[0.08, 0.37]	0.11	.003
		Study Quality ⇒ Change in Feelings	0.07	0.05	[-0.02, 0.18]	0.07	.14
	Direct	Condition ⇒ Change in Feelings	0.00	0.09	[-0.18, 0.18]	0.00	.98
	Total	Condition ⇒ Change in Feelings	0.03	0.11	[-0.19, 0.25]	0.01	.78

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S3.7. *Moderated mediation estimates of condition predicting change in beliefs about bias by prior bias beliefs in the conservative-friendly materials of Study 3.*

Prior Bias Beliefs Group	Type	Effect	Estimate	SE	95% CI	*β*	*p*
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	-0.08	0.03	[-0.15, 0.02]	-0.03	.01
		Condition ⇒ Study Quality	0.25	0.09	[0.07, 0.44]	0.13	.009
		Study Quality ⇒ Change in Beliefs	-0.33	0.05	[-0.43, -0.25]	-0.23	< .001
	Direct	Condition ⇒ Change in Beliefs	0.01	0.11	[-0.20, 0.22]	0.00	.95
	Total	Condition ⇒ Change in Beliefs	-0.05	0.12	[-0.28, 0.18]	-0.02	.65
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	0.13	0.03	[0.06, 0.20]	0.05	< .001
		Condition ⇒ Study Quality	-0.28	0.07	[-0.41, -0.16]	-0.15	< .001
		Study Quality ⇒ Change in Beliefs	-0.46	0.05	[-0.56, -0.37]	-0.32	< .001
	Direct	Condition ⇒ Change in Beliefs	-0.18	0.08	[-0.34, -0.01]	-0.06	.036
	Total	Condition ⇒ Change in Beliefs	0.02	0.08	[-0.14, 0.18]	0.01	.78
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	0.48	0.08	[0.32, 0.64]	0.17	< .001
		Condition ⇒ Study Quality	-0.82	0.09	[-0.99, -0.64]	-0.43	< .001
		Study Quality ⇒ Change in Beliefs	-0.59	0.07	[-0.73, -0.45]	-0.39	< .001
	Direct	Condition ⇒ Change in Beliefs	-0.36	0.14	[-0.64, -0.09]	-0.13	.01
	Total	Condition ⇒ Change in Beliefs	0.10	0.12	[-0.13, 0.33]	0.04	.40

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.

Table S3.8. *Moderated mediation estimates of condition predicting change in beliefs about bias by prior bias beliefs in the liberal-friendly materials of Study 3.*

Prior Bias Beliefs Group	Type	Effect	Estimate	SE	95% CI	β	p
Conservative-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	-0.31	0.07	[-0.45, -0.17]	-0.12	< .001
		Condition ⇒ Study Quality	-0.55	0.11	[-0.75, -0.34]	-0.26	< .001
		Study Quality ⇒ Change in Beliefs	0.57	0.06	[0.46, 0.68]	0.47	< .001
	Direct	Condition ⇒ Change in Beliefs	0.23	0.13	[-0.02, 0.47]	0.09	.065
	Total	Condition ⇒ Change in Beliefs	-0.10	0.11	[-0.32, 0.11]	-0.04	.36
Average	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	-0.10	0.03	[-0.16, -0.04]	-0.04	.002
		Condition ⇒ Study Quality	-0.20	0.06	[-0.33, -0.08]	-0.10	.001
		Study Quality ⇒ Change in Beliefs	0.48	0.04	[0.40, 0.55]	0.40	< .001
	Direct	Condition ⇒ Change in Beliefs	0.06	0.07	[-0.09, 0.20]	0.02	.44
	Total	Condition ⇒ Change in Beliefs	-0.08	0.08	[-0.23, 0.08]	-0.03	.33
Liberal-leaning	Indirect	Condition ⇒ Study Quality ⇒ Change in Beliefs	0.05	0.03	[-0.01, 0.12]	0.02	.11
		Condition ⇒ Study Quality	0.14	0.08	[-0.02, 0.30]	0.07	.09
		Study Quality ⇒ Change in Beliefs	0.38	0.05	[0.29, 0.48]	0.33	< .001
	Direct	Condition ⇒ Change in Beliefs	-0.12	0.09	[-0.29, 0.05]	-0.05	.17
	Total	Condition ⇒ Change in Beliefs	-0.05	0.11	[-0.27, 0.17]	-0.02	.65

Note. Condition was dummy-coded (0 = Blinded). Conservative-leaning = M -1SD, Liberal-leaning = M + 1SD. Confidence intervals computed with 1000 bootstrap replications.