UNIVERSITY OF CALIFORNIA SAN DIEGO

How the Consideration of Negative Stimuli Shapes Individuals' Preferences and Behavior

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Management

by

Giulia Maimone

Committee in charge:

Professor Ayelet Gneezy, Co-Chair
Professor Uma R. Karmarkar, Co-Chair
Professor On Amir
Professor Uri Gneezy
Professor Joachim Vosgerau
Professor Piotr Winkielman

2022

The Dissertation of Giulia Maimone is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my beloved family and friends, who have painfully accepted my absence to let me follow my dream, to my adored nieces, Alessandra and Beatrice, who have been my constant light and motivation, and to my loving and supportive husband, Patrick, who has been by my side for every second of this wonderful and tumultuous journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the invaluable guidance and support of so many professors, coauthors, peers, friends, and family. I will be forever grateful for and humbled by the incredible opportunity that has been given to me.

First and foremost, I would like to thank my amazing advisors, Ayelet Gneezy and Uma Karmarkar. The first time I entered the United States of America was the day I started my PhD. I took a shuttle to school and thought it would be a good idea to go introduce myself to the professors in my department. I arrived on the third floor at Rady, in the professors' office area, and started knocking on their doors. I clearly remember Ayelet opening her door that day and inviting me in. We started chatting about research, and I immediately fell in love with her curiosity and open-mindedness. By the end of the hour, she was my advisor. I truly feel like the luckiest woman on Earth to have had her by my side for the whole length of my PhD, and—I am sure—beyond it. She is not only an incredibly talented, smart, creative, and hard-working scholar, but also one of the most caring, involved, loving, special people I know. I will cherish her friendship, mentorship, and advice forever. I would not be the person and the researcher I am today without her. I thank her from the bottom of heart for always encouraging me to study the questions I was truly passionate about and for helping me discover my identity as a researcher.

Uma became my advisor in a critical stage of my PhD journey, my third year, when I was trying to find my way and right before the COVID pandemic hit. I will always be grateful for her support and mentorship during such a confusing time. I was always struck by Uma's unique way to think about research. I have always perceived her as a scientist rather than a researcher, and I

have admired her so much for that status. Looking up to her and "doing science" with her had an enormous impact on my own view of the field, of myself, of my work, and of the responsibilities that come with it. Uma challenged me to push my boundaries and leave my comfort zone. She didn't simply make me a better researcher; she made me a scientist, too. I am humbled by the awareness that this perspective will tremendously impact every decision I will make in my work and will color my vision of the world.

On Amir has also played a leading role in my time at Rady and my development as a researcher. I thank On for his generosity with his time and knowledge. I will always remember our conversations about research and ideas, but also music and Marvel movies! On is one of those impressive people who seem to (or, more likely, actually do) know everything about everything. I am very lucky to be able to call On a co-author. Working with him is a unique experience, because he is able to infuse a sense of calm and confidence in you with his totally logical and rational approach to research. I am so thankful for the trust and optimism I have always felt while leaving his office. I would go in with a problem and come out with either a solution or the reassurance that it wasn't a problem after all. On taught me that failure doesn't exist until there is something to learn from it. I am, and always will be, grateful for his teachings, confident guidance, advice, and friendship.

I owe my deepest gratitude and a special thank you to Joachim Vosgerau. When I met Joachim, in March 2016, I had just graduated with my Master of Science from Bocconi University and knew nothing about consumer behavior research. Joachim not only hired me as his research assistant and lab manager, but also, in the following year and a half, spent an impressive amount of time talking to me about research, answering every question I asked, teaching me (I will never forget him drawing graphs on the windows of his office to explain to

me the regression to the mean or the issues of self-selection), and working with me on my own research ideas. I will never be able to convey in words how unconditionally grateful I am for the time and effort Joachim dedicated to me in those months, and for having believed in me and my potential. Without his teaching and support, I would not be here today writing this acknowledgement. As if his guidance during my time as his research assistant weren't enough, Joachim continued to support me and collaborate with me for the entire length of my PhD. Joachim is a certainty for me in both my professional and personal life, as we developed a close friendship over the years. I am so proud that Joachim had such a profound influence on the researcher and the person I am today. Thank you!

I am grateful to my other committee members as well. During the first two years of my PhD, I worked as Uri Gneezy's lab manager at Rady. Uri is an incredibly fair, kind, and supportive person, and so he was as a boss. Working for him has been an incredible learning experience and a true honor. I was also very lucky to take three of Uri's classes. Hearing him talking about research and learning how he thinks about research is simply amazing, and my time with him has been a huge privilege. Throughout my PhD, I would regularly stop by his office to chat about research and life, and I will forever be grateful for his unconditional support and generosity with his time. I also would like to thank Piotr Winkielman. I met Piotr when I took his social psychology class early in my PhD, and I immediately appreciated his extraordinarily kind and friendly personality, and his deep knowledge and perspective of social psychology. Later on, I found out Piotr is also incredibly curious and knowledgeable about many other topics—among which is Italian history! I had the most interesting conversations with him, and it was so humbling to see such a successful scholar continue to be curious and eager to learn new things. I am very grateful to have him on my committee.

I would additionally like to thank several amazing coauthors who are not on my committee. During my time at Rady, I have been extremely lucky to work with Craig McKenzie, whom I met when I took two of his Judgment and Decision-Making classes. The first time we spoke, it was immediately clear to me that we shared an interest in the same types of questions, and I was absolutely amazed by his independent thinking, his approach to research, and his kindness. One of my favorite things about Craig is that—despite our very obvious disparity in terms of rank and experience—he always made me feel like a peer, and not once did I feel like a student talking to a professor. Several of our conversations led naturally to research projects that I thoroughly enjoyed working on. Craig has played a huge role in my development as a researcher. I am beyond grateful for his friendship and constant support, and I am looking forward to many more collaborations to come! I also would like to thank Gil Appel. Ayelet introduced me to Gil in my third year, and we immediately became not only enthusiastic co-authors, but also close friends. Gil is one of the most generous people I know. He has served as a mentor and taught me so much about successful interdisciplinary collaborations. I will be forever grateful for the countless hours we spent working together on our projects. Last but not least, another great researcher I was lucky enough to work with—even though for too short a time—is Tom Meyvis. Tom was not only an extremely brilliant and generous scholar, but also a wonderful person. I am deeply grateful I had the chance to work with Tom on an exciting project, to receive his advice, and to experience his innate kindness.

I also would like to thank all the current and former Rady faculty and postdocs who provided encouragement, feedback, and guidance during my time at Rady: Wendy Liu, Rachel Gershon, Ken Wilbur, Karsten Hansen, Kanishka Misra, Vincent Nijs, Robert Sanders, Michael Meyer, Chris Oveis, Yuval Rottenstreich, Elizabeth Campbell, Tage Rai, Sally Sadoff, Marta

Serra-Garcia, Anya Samek, Charlie Sprenger, Evan Weingarten, and Jerry Grimes. Their feedback made my research and presentations stronger and more interdisciplinary. I am grateful for their time, knowledge, and support. I owe a special thank you to the fantastic Pam Smith, who has always been so supportive, helpful, and available to chat, give constructive feedback, and provide us students a platform to share and discuss our work and ideas, her legendary lab (that changed name many times, but for me it will always remain Pam's lab!), and Alison Bloomfield Meyer, who helped me so much to develop my presentation skills and to find my own voice. I am also extremely grateful for the Rady IT and administrative staff. Their work and support have been crucial for my research, teaching, and learning experience at Rady. I am particularly grateful for the irreplaceable Patricia, who—beyond having provided impeccable support for pretty much any activity I was involved in—brightened up my days with her kindness and warmth. I am also so thankful to Hillary, the best PhD coordinator I could have ever hoped for. She is able to make the impossible possible. Additionally, I thank the legendary Erwin who keeps Rady going and who has been like a guardian angel for anything that happened at Rady in all these years. I truly cannot imagine my time at Rady without them. Finally, I am so grateful for Dean Ordóñez, who has always been such a supportive leader for all of us.

Many amazing colleagues and friends have also been by my side during this journey. First and foremost, I am deeply grateful for Ariel. Ariel and I shared a cubicle first and then an office for my entire time at Rady. My PhD experience would not have been the same without his friendship and support and our proofreading of each other's emails. I also owe the deepest thank you to Seung Hyun Kim, Jean Zhang, Michelle Kim, Meenakshi Balakrishna, Olivia Jurkiewicz, Anindo Sarkar, Gal Smitizsky, Heather Romero-Kornblum, Shuang Wu, Paul Wynns, Brianna Chew, Katie Hillegass, Carolina Raffaelli, Amanda Nachman, John-Henry Pezzuto, Arathy

Puthillam, and Gabriella Wong for their friendship and for having shared so much with me over the years. I am also grateful for the amazing former PhD students I was lucky enough to cross paths and build friendships with during my time at Rady. Kristen Duke, Allie Lieberman, Jessica Kim, Min Zhang, Yidan Yin, Yumeng Gu, and Lim Leong will always be inspiring role models for me! I am also thankful to the current and former Rady Behavioral Lab managers, Tiss Doci and Jyoti Jhita, and all the lab's research assistants who worked tirelessly to run experiments and make sure data were always collected properly.

The academic colleagues who have become close friends are also owed my gratitude. Elisa Solinas, Maria Giulia Trupia, Martina Cossu, Mohamed Hussein, Jimin Nam, Anna Tari, Jinwoo Kim, Jin Kim, Samuel Levy, Nofar Duani, Maria Langlois, Alexander Park, Graham Overton, Rodrigo Dias, David Dolifka, and many more made conferences the time of the year to look forward to.

Last but not least, words will never be enough to express the gratitude and love I have for my family. My parents, Elisa and Fabio, made so many sacrifices to allow me to be where I am today, dealing with my absence, and providing me love and support every day of my life. I am so grateful and proud to be their daughter. My sister and brother, Claudia and Federico, also play such an enormous role in my life, I will forever be grateful to know that, whatever happens, I will always have them in my corner. My amazing and adored nieces, Alessandra and Beatrice, are the reason and motivation behind every one of my choices. I love and miss them more than words can say. I am grateful for my behind-the-scenes strength, my forever best friends, Anna, Marta, Elisa, Nathalie, Chiara, and Donatella. I could not have done this without you cheering me on every step of the way. I also feel beyond lucky for my dear friends Luisa and Alessandro, who have made San Diego feel like home whenever I missed Italy. Additionally, I would like to

thank my amazing parents-in-law, Margie and Mike, who have welcomed me like a daughter. With their love and warmth, they have made the distance from my family much more tolerable, and I will be forever grateful for them. Finally, I undoubtedly owe the biggest thank you to my wonderful husband, Patrick. Meeting him at the math bootcamp before starting our PhD programs has been the blessing of my life. I feel so lucky to have met him; I could not imagine my life without him. Beyond being just a wonderful human being, Patrick is an incredibly curious, smart, hard-working, and creative scholar. Going through the PhD journey together was not only a huge source of motivation, but also the reason why I reached the end of it. He has been my strength, my rock, my safe harbor. Thank you from the bottom of my heart!

Chapter 1, in full, has been submitted for publication of the material. Maimone, Giulia, Gil Appel, Craig R. M. McKenzie and Ayelet Gneezy. The dissertation author is the primary investigator and author of this paper.

Chapter 2, in full, has been submitted for publication of the material. Maimone, Giulia, Joachim Vosgerau and Ayelet Gneezy. The dissertation author is the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material. Maimone, Giulia, Uma R. Karmarkar and On Amir. The dissertation author is the primary investigator and author of this paper.

VITA

2013	Bachelor of Science in Management, Bocconi University
2015	Master of Science in Management, Bocconi University
2016-2017	Research Assistant and Lab Manager, Bocconi University
2017-2019	Lab Manager, Rady School of Management, University of California San Diego
2019-2022	Teaching Assistant, Rady School of Management, University of California San Diego
2020-2022	Teaching Assistant, School of Global Policy and Strategy, University of California San Diego
2022	Doctor of Philosophy in Management, Rady School of Management, University of California San Diego

ABSTRACT OF THE DISSERTATION

How the Consideration of Negative Stimuli Shapes Individuals' Preferences and Behavior

by

Giulia Maimone

Doctor of Philosophy in Management

University of California San Diego, 2022

Professor Ayelet Gneezy, Co-Chair
Professor Uma R. Karmarkar, Co-Chair

This dissertation contains three papers that together shed light on how and why the consideration of negative stimuli (i.e., *information*, *outcomes*, *language*) shapes individuals' preferences and behavior.

Chapter 1 investigates how the consideration of *negative information* about a scholar—unrelated to their work—affects other scholars' citing behavior. Academics use citations to acknowledge the contribution of past work and promote scientific advancement. However, analyzing citation data of 32,025 publications spanning 18 academic fields, we find evidence suggesting citations may also serve as a currency to reward and punish scientists' morality. We find that, relative to controls, the citation rates of scholars accused of sexual misconduct decrease after the accusations become public. Interestingly, this citation penalty is *larger* than the one incurred by scientists accused of scientific fraud. Our findings suggest that, in addition to serving the purpose of maintaining intellectual integrity and promoting scientific advancement, citation decisions are also driven by scholars' attitudes toward the publications' authors.

Chapter 2 demonstrates how and why the consideration of *negative outcomes* impacts people's attributional preferences. There are two streams of literature that address attributional preferences: self-determination and self-serving preferences. While these two theories make the same prediction for individuals' attributional preferences over positive outcomes, they make competing predictions for attributional preferences over negative outcomes. Self-determination maintains that people prefer to have agency over negative outcomes. Self-serving preferences, in contrast, stipulate that people prefer to concede agency over negative outcomes. In eight preregistered experiments (N = 3,946), we reconcile these seemingly inconsistent attributional preferences over negative outcomes. First, we test these competing predictions and find that—consistent with self-determination—people would rather "own" their negative outcomes than externally attribute them. Overplacement (people's belief that they perform better than others) and the impact bias (the belief that "owned" negative outcomes hurt less than when they are caused by oneself) cannot explain this preference. Instead, we find that reducing the saliency of

agency moderates the preference for agency over negative outcomes. More interestingly, we find that sharing agency reverses attributional preferences: while people prefer assuming agency over negative outcomes when these are exclusively caused by a sole agent (either themselves or somebody else), they prefer attributing agency to others when negative outcomes are jointly caused by multiple agents (both themselves and somebody else).

Finally, chapter 3 examines how different cognitive processes elicited by *negative language* affect a message's efficacy. Across five preregistered field and lab experiments (N = 22,024), we demonstrate when and how using an easily reversible (i.e., bi-polar) word in a statement, rather than a non-reversible one with the same meaning, engages different cognitive processes and leads to different outcomes. In particular, when a statement containing a bi-polar word is processed as a negation (i.e., opposing a claim rather than affirming it), a slower more elaborate cognitive process occurs. We show that this results in lower judgment confidence, and a lower likelihood to act on the message. In addition, we find that this more elaborative process also leads to weaker attitudes towards the message source. Our findings advance consumer theories by shedding light on the ways in which linguistic elements of communication impact judgments and real-world behaviors. They additionally offer practical persuasive messaging strategies for those engaged in a range of marketing and policy communications.

Chapter 1.

CITATION PENALTIES FOLLOWING SEXUAL MISCONDUCT VERSUS SCIENTIFIC FRAUD ALLEGATIONS

Giulia Maimone, Gil Appel, Craig R. M. McKenzie, Ayelet Gneezy

Rady School of Management, University of California, San Diego, La Jolla, CA, 92093, USA

ABSTRACT

Academics use citations to acknowledge the contribution of past work and promote scientific advancement. However, analyzing citation data of 32,025 publications spanning 18 academic fields, we find evidence suggesting citations may also serve as a currency to reward and punish scientists' morality. We find that, relative to controls, the citation rates of scholars accused of sexual misconduct decrease after the accusations become public. Interestingly, this citation penalty is *larger* than the one incurred by scientists accused of scientific fraud. Our findings suggest that, in addition to serving the purpose of maintaining intellectual integrity and promoting scientific advancement, citation decisions are also driven by scholars' attitudes toward the publications' authors.

Powerful social movements such as #MeToo have increased the public's awareness of the pervasiveness of sexual transgression across industries (*1*) and demanded that perpetrators be held accountable (*2*, *3*). Academia has not been spared (*4, 5*). For example, a 2021 report estimates 20% of female and 6%–8% of male undergraduates were victims of sexual misconduct during their college life (*6*). For the victims, the consequences of sexual violence are devastating and may include emotional, physical, and professional outcomes (7). Academic institutions abide by Title IX of the Education Amendments Act of 1972—prohibiting sex-based discrimination under any education program—to regulate processes and sanctions for dealing with transgressors and deterring such behaviors. In addition, individuals may express disapproval or punish deviants in informal ways, such as sharing information on social media (*8*).

In this article, we consider another potential course of action available for academics wishing to express their disapproval of a peer's misconduct. In particular, we investigate the effect of sexual-misconduct allegations on the citation rates of the alleged perpetrators' work. The purpose of citations is to promote scientific advancement and acknowledge the contribution of past research (*9, 10*). Consequently, across disciplines, the number of citations a scholar has provides a measure of the quality and impact of their work (*11–14*). Interestingly, evidence suggests scientists' decision to cite an article is sometimes driven by non-scientific reasons. For example, researchers are more likely to cite their friends' articles, in part to help their friends and in part to help themselves, because they are more likely to be cited in return (*15*, *16*). Whether scholars might also choose *not* to cite research in order to *hurt* their peers is unclear.

If scholars use citations exclusively for scientific purposes, allegations of sexual misconduct should not influence the citation rate of the accused's research, because their transgression does not implicate the relevance or quality of their work. At the same time, if

researchers' citation decisions were also sensitive to non-scientific factors, the citation rate of the accused might be negatively affected. Specifically, scholars might avoid citing the research of a colleague accused of sexual misconduct to signal disapproval, distance themselves from the alleged perpetrator, or punish them, producing a negative effect of sexual-misconduct allegations on the accused's citation rates.

To determine whether citations of scholars accused of sexual misconduct decrease, we compare their citation rates with those of control scholars (see Methods). A slower increase (or faster decrease) in citation rates after the allegations became public for scholars accused of sexual misconduct would suggest a citation penalty. To further contextualize the size of any observed citation penalty for scholars accused of sexual misconduct, we compare these scholars' citation rates with those of researchers accused of scientific fraud (e.g., data fabrication, falsification, and plagiarism; *17, 18*). Unlike sexual misconduct, scientific fraud necessarily implicates the concerned research, invalidating its claims (*19*), and may generate concerns about the integrity of the entire portfolio of the accused researcher. From this perspective, a reasonable expectation is a *smaller* citation penalty for scholars accused of sexual misconduct than for those accused of scientific fraud.

Consistent with the hypothesis that scholars may use citations to hurt their peers, our analyses (N = 32,025 publications) reveal a significant citation penalty for scholars accused of sexual misconduct. Importantly, although scientific fraud implicates the accused's scientific contribution whereas sexual misconduct doesn't, we also find the citation penalty for research published by scholars accused of sexual misconduct is substantially *larger* than that observed for research published by scholars accused of scientific fraud.

Our data consist of academic citations for 32,025 publications (peer-reviewed articles and scholarly books; collected from the Web of Science platform) across 18 disciplines. Of these, 6,012 publications are authored by thirty scholars accused of sexual or scientific misconduct (split evenly), and 26,087 publications are authored by closely matched controls (see Supplementary Information for details). All accused scholars included in our sample were active researchers in the natural or social sciences, having a minimum of 200 citations overall, and were accused of misconduct in 2017 or earlier.

To ensure that our findings are not biased by differential awareness of the two misconduct types, we restrict our sample to accused scholars whose accusations received similar media attention. First, we searched for online news reports on sexual and scientific misconduct in academia to identify relevant cases for our analyses. Next, to validate the integrity of our selection process, we checked each case against the following databases: the Academic Sexual Misconduct Database, which lists sexual-misconduct allegations involving faculty and other university employees (22), Retraction Watch—a blog that reports on retractions of scientific papers (23), and Wikipedia's List of Scientific Misconduct Incidents (24). Although this requirement limits the number of misconduct cases accounted for, the total number of publications in our data (N = 32,025) provides sufficient power to detect significant and reliable results.

For every accused scholar, we matched five researchers not accused of any misconduct. Control researchers were matched to their respective accused scholar by their similarity along the following dimensions: (a) research discipline, (b) research topics, (c) seniority, (d) a comparable university, and (e) a comparable number of total citations. Following the collection process, we excluded eight control researchers (4 sexual and 4 scientific) for having at least one retracted

paper.[1] Spanning a maximum of 13 years,[2,3] our data account for citations of 6,012 publications authored by 30 accused scholars and 26,087 publications authored by 142 closely matched control scholars, totaling 294,026 observations. All analyses control for publication year, total citations per paper, number of authors, the scholar's academic discipline, rank, gender, and the year the accusations became public (see Supplementary Material for robustness checks).

First, we compare citation rates of researchers accused of sexual misconduct, researchers accused of scientific misconduct, and their respective controls. A difference-in-difference-in-differences analysis comparing citation rates of these four groups before and after news of the allegations broke shows the average yearly citations per publication of the two control groups increased ($b_{Scientific}$ = 0.67, t(293,993) = 6.89, p < .001; b_{Sexual} = 0.95, t(293,993) = 10.28, p < .001). Citation rates of scholars accused of scientific fraud remained unchanged (b = -0.13, t(293,993) = -0.70, p = .483), and those of scholars accused of sexual misconduct decreased (b = -0.74, t(293,993) = -3.87, p < .001).[4] Next, we ran the same regression controlling for the natural increasing trend in citations over time (*25*, *26*)—depicted in Figure 1.1.[5] Thus, to the extent that a line is flat, citations increased normally—that is, with no penalty. The results show the citation rates of both control groups are flat ($b_{Scientific}$ = -0.07, t(293,992) = -0.59, p = .556; b_{Sexual} = 0.21, t(293,992) = 1.72, p = .086), indicating no penalty. In contrast, citations rates of scholars accused of either misconduct type decreased ($b_{Scientific}$ = -0.86, t(293,992) = -4.18, p < .001; b_{Sexual} = -1.49, t(293,993) = -7.16, p < .001), indicating a citation penalty. Importantly, our analysis reveals a

[1] Results are robust to including all 150 controls; see Supplementary Material.
[2] We account for 10 years before (pre-accusation) and 3 years after (post-accusation) the accusations became public. Later publications might have fewer observations (i.e., years), depending on when they were published.
[3] Results are robust to different specifications of the number of years (i.e., 5, 8, 12, and 15) in the pre-accusation period (see Supplementary Material).
[4] See Supplementary Material (Figure 6).
[5] Regression estimates and p-values of interaction terms are similar regardless of whether we control for trend; see Supplementary Material.

greater citation penalty for scholars accused of sexual misconduct than for scholars accused of scientific misconduct, qualified by a significant three-way interaction (b = -0.92, t(293,992) = -3.06, p = .002)[6] of misconduct type (sexual vs. scientific), accusation (accused vs. control), and time (pre- vs. post-accusation). This indicates that the difference between the citation rates of scholars accused of sexual misconduct and their own controls is larger than the difference in citations between scholars accused of scientific fraud and their controls. Finally, consistent with a larger citation penalty for scholars accused of sexual misconduct, a direct comparison of citation rates of the two accused groups before and after the allegations became public reveals a significant interaction. Specifically, the citations of scholars accused of sexual misconduct decreased *more* than those of scholars accused of scientific fraud (b = -0.63, t(293,992) = -2.35, p = .019).

We consider two unobserved factors that could explain why research by scholars accused of scientific misconduct incurs a relatively small citation penalty. Non-retracted publications of scholars accused of scientific fraud could receive a boost if, for example, individuals interpret the absence of retraction as suggesting a publication has been cleared. Alternatively, retracted publications might receive a boost if researchers cite a retracted article when referring to its shortcomings. To test these possibilities, we regressed average yearly citations of researchers accused of scientific misconduct on retraction status (yes vs. no), time (pre- vs. post-accusation), and their interaction. Ruling out both alternative explanations, our analysis reveals scientific-fraud accusations hurt the citations of both retracted (b = -3.20, t(28,568) = -4.38, p < .001) and non-retracted (b = -1.80, t(28,568) = -8.24, p < .001) publications. Citation rates of

[6] We replicate this result using a variety of robustness checks; see Supplementary Material.

retracted publications decreased marginally more than those of the non-retracted publications (b = -1.40, t(28,568) = -1.93, p = .054; see Figure 1.2).

Finally, researchers may simply be more likely to know about scholars involved in sexual—versus scientific—misconduct, for example, if sexual misconduct gain more traction in the media and informal platforms. Although we cannot rule out this possibility, we selected both the sexual- and the scientific-misconduct cases based on information available online, minimizing such concern in our data. In addition, while sexual-misconduct allegations are sometimes protected by privacy tools (e.g., non-disclosure agreements), scientific-fraud allegations are not. In fact, Google Scholar, Web of Science, and journal websites all clearly mark retracted articles.

Despite the direct relationship between scientific fraud (vs. sexual misconduct), the integrity of the alleged offender's research, and citations, our analysis shows scholars accused of sexual misconduct incur a larger citation penalty than scholars accused of scientific fraud. These findings complement other research showing non-replicable papers are cited more than replicable ones (*27*) and that false positives in science tend to persist rather than self-correct (*28–32*), despite the availability of tools designed to address these challenges (*33–35*).

The pattern observed in our data could be driven by insufficient sensitivity to scientific misconduct, oversensitivity to sexual misconduct, or both. To test whether differential sensitivity is at play, we presented individuals (N = 231)[7] with definitions and examples of both scientific and sexual misconduct in academia. Participants indicated which of the two types of misconduct they thought was (a) more deserving of punishment, (b) more disgusting, and (c) worse than the

[7] Two hundred fifty-one Amazon Mechanical Turk workers were recruited; we excluded twenty for failing our attention check.

other.[8] The overwhelming majority of participants deemed sexual misconduct in academia as more deserving of punishment (76.2%, $\chi^2(1) = 63.38$, $p < .001$), more disgusting (90.5%, $\chi^2(1) = 151.40$, $p < .001$), and worse (75.8%, $\chi^2(1) = 61.30$, $p < .001$) than scientific misconduct. Participants' gender had no effect on any of our measures (ps > .40). Researchers might similarly have greater negative affective reactions to sexual misconduct and be less likely to cite work published by an alleged sexual offender in order to punish them (*36*, *37*), to signal they condemn their actions (*38*), or simply because they exhibit avoidance behavior toward a stigmatized misconduct type (*39*). Note our investigation focuses on the *effect* of sexual misconduct on citation rates of the accused. In referring to the decision not to cite an alleged perpetrator's work as punishing or hurting them, we are merely describing the outcome of that decision, not the citing scholar's intent. Another factor that might contribute to our findings could be people's need to take action in light of the evident inadequacy of various policies (e.g., Title IX) and of the justice system to keep sexual offenders accountable and deter future transgressions (*8*, *40*).

To the best of our knowledge, our paper is the first to offer a systematic comparison of the ramifications of sexual misconduct and scientific misconduct on the citations of alleged perpetrators. Our finding that the citation penalty for sexual-misconduct allegations is larger than the citation penalty for scientific misconduct provides additional evidence that citation decisions are sensitive to factors unrelated to a publication's scientific merit (see *15*). Specifically, it suggests that, in addition to serving the purpose of maintaining intellectual honesty and promoting scientific advancement, citations also serve as a currency to benefit and hurt other scientists. These finding contribute to recent discussions concerning research practices (*33*), editorial decisions (*10*), and the integrity of scientific research (*21*, *38*).

[8] See Supplementary Material and https://osf.io/ycazs/?view_only=bfb1080d756146ba89d21a7ed3daeacf for details.

ACKNOWLEDGEMENTS

The authors thank Uri Gneezy, Leif Nelson, and Shlomi Sher for helpful comments.

Chapter 1, in full, has been submitted for publication of the material. Maimone, Giulia, Gil Appel, Craig R. M. McKenzie and Ayelet Gneezy. The dissertation author is the primary investigator and author of this paper.

DATA AND MATERIALS AVAILABILITY

All data, code, and materials used in the analyses are available on OSF at the following link: https://osf.io/ycazs/?view_only=bfb1080d756146ba89d21a7ed3daeacf

FIGURES

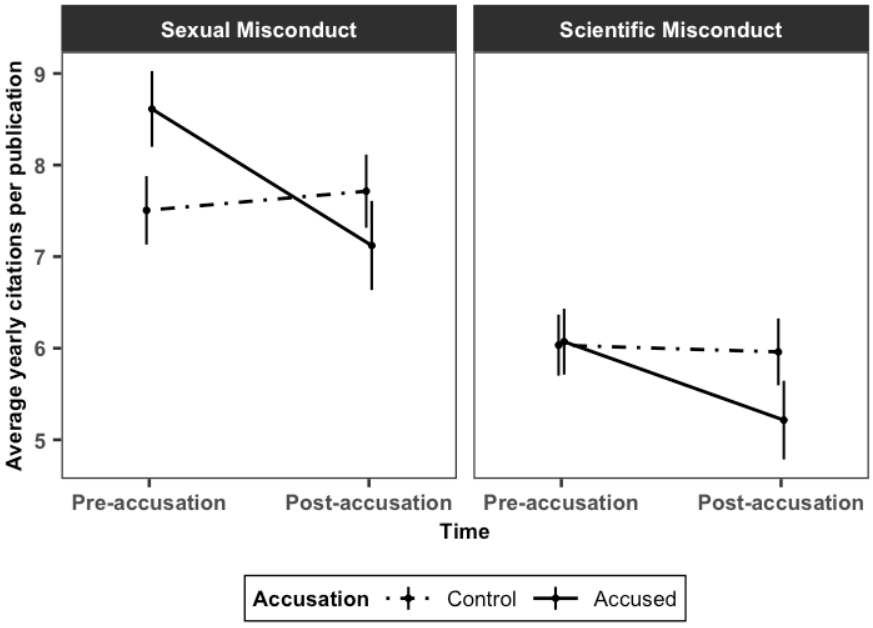


Figure 1.1. Average yearly citations per publication by accusation, misconduct type, and time

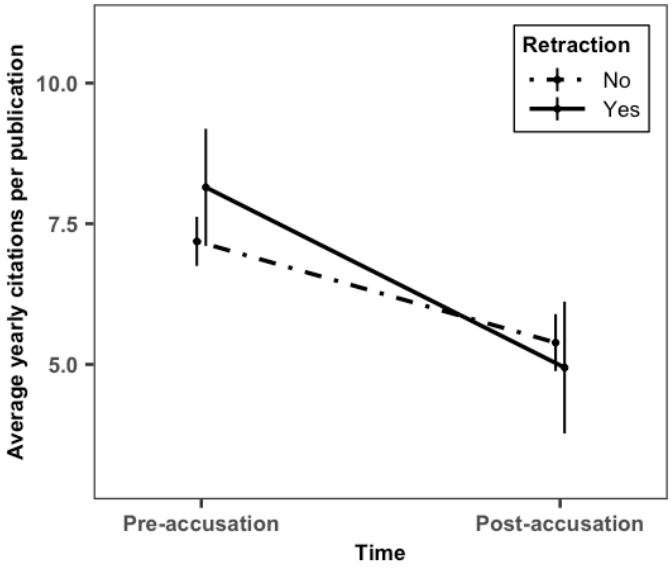


Figure 1.2. Average yearly citations (of scholars accused of scientific fraud) per publication by retraction status and time

APPENDIX

CONTROLS' SELECTION

We collected 5 controls for every accused scholar. We matched control and accused scholars according to the following criteria:

1. Same field as the accused scholar
2. Similar research topic as the accused scholar
3. Similar rank as accused scholar
4. Similar overall number of citations as the accused scholar
5. Comparable university (given field)
6. No public accusations

DATA EXCLUSIONS

In addition to excluding 8 controls who had at least one retracted paper, we excluded the following publications:

1. Publications coauthored by multiple accused scholars in our sample
2. Publications coauthored by an accused scholars in our sample and one of their own controls
3. Publications that are retraction reports (i.e., articles that announce a certain publication will be retracted)

MAIN ANALYSIS (CONTROLLING FOR TIME TREND) ADDITIONAL MODEL: COLLAPSING ACCUSED SCHOLARS IN A SINGLE GROUP

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.08, t(293,996) = 0.72, p = .470), those of the accused scholars decreased substantially more after the accusations became public (b = -1.25, t(293,996) = -8.30, p < .001).

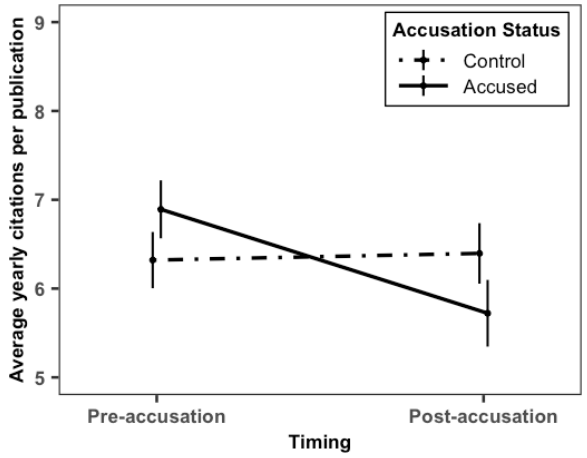


Fig. 1. Change in average yearly citations per publication by accusation status from before to after the misconduct accusations became public.

REGRESSION ROBUSTNESS CHECKS

In this section we test our findings from Model 1, Model 2, and Model 3 across different regression specifications. Specifically, we ran each of our models in four different ways:

- Not controlling for the natural trend of citations (i.e., No time trend)
- Not controlling for any covariate (i.e., No covariates)
- Having the 172 scholars as random effects (i.e., Scholars RE)
- Having the disciplines our scholars work in as random effects (i.e., Field RE)

Covariates' description

- Publication Year = Year in which the publication was published: *number*
- Publication's Citations = Total number of citations of the publication: *number*
- Publication's # of Authors = Number of authors of the publication: *number*
- Scholar's Field = Scholar's discipline: *Astronomy, Astrophysics, Biochemistry, Bioinformatics, Cellular Biology, Chemistry, Economics, Electrical Engineering, Geology, Marketing, Medicine, Microbiology, Molecular biology, Neuroscience, Political Science, Psychology, Psychology/Neuro*[9]*, Statistics*
- Scholar's Rank = Scholar's rank at the time of data collection or last position held in academia: *Associate Professor, Full Professor, Emeritus Professor*
- Scholar's Gender = Scholar's gender: *Male, Female*
- Accusations Year = Year in which the accusations to the accused scholar of reference became public: *number*
- Time Trend = the considered year and the year in which the accusations became public: *-10*, *-9*, *-8, -7, -6, -5, -4*, *-3*, *-2*, *-1*, *1*, *2*, *3*
- Scholar = Author of the publication we are considering either as accused of sexual misconduct, accused of scientific misconduct, or as a control

[9] Scholar publishes in Neuroscience but is/was hired in a Psychology department.

Regression Tables

Model 1: 2(Accusation Status: Control vs. Accused) x 2(Timing: Pre- vs. Post-Accusation)

	Average Yearly Citations Per Publication			
	No time trend (1)	No covariates (2)	Scholars RE (3)	Field RE (4)
Post-accusation	0.808***	1.107***	0.074	0.076
	(0.067)	(0.129)	(0.104)	(0.104)
Accused	0.569***	-0.316*	0.471	0.574***
	(0.087)	(0.165)	(0.328)	(0.087)
Post-accusation:Accused	-1.249***	-1.530***	-1.170***	-1.246***
	(0.150)	(0.295)	(0.150)	(0.150)
Covariates	Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year		Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year Time Trend	Publication Year Publication's Citations Publication's # Authors Scholar's Rank Scholar's Gender Accusations Year Time Trend
Observations	294,026	294,026	294,026	294,026
Random Effect			Scholar	Field
RE Groups			172	18
RE Variance			221.69	223.72
R2	0.742	0.000	0.742	0.741
Adjusted R2	0.742	0.000	0.745	0.742
Residual Std. Error	15.000 (df = 293,997)	29.400 (df = 294,022)	14.89	14.96

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 2. Regression table for Model 1 – Regression robustness checks[10].

[10] Location of models' descriptions – No time trend (1): Fig. 5; No covariates (2): Fig. 8; Scholas RE (3): Fig. 11; Field RE (4): Fig. 14.

Model 2: 2(Accusation Status: Control vs. Accused) x 2(Timing: Pre- vs. Post-Accusation) x 2(Misconduct Type: Scientific vs. Sexual)

	Average Yearly Citations Per Publication			
	No time trend (1)	No covariates (2)	Scholars RE (3)	Field RE (4)
Post-accusation	0.666***	1.367***	-0.111	-0.075
	(0.097)	(0.188)	(0.125)	(0.125)
Accused	0.039	0.530**	0.051	0.035
	(0.129)	(0.237)	(0.469)	(0.129)
Sexual	1.469***	0.825***	1.872***	1.408***
	(0.149)	(0.147)	(0.541)	(0.146)
Post-accusation:Accused	-0.799***	-1.847***	-0.710***	-0.781***
	(0.212)	(0.417)	(0.212)	(0.212)
Accused:Sexual	1.062***	-1.634***	0.685	1.074***
	(0.176)	(0.330)	(0.642)	(0.176)
Post-accusation:Sexual	0.279**	-0.503*	0.349***	0.284**
	(0.132)	(0.259)	(0.132)	(0.132)
Post-accusation:Accused:Sexual	-0.890***	0.569	-0.914***	-0.920***
	(0.300)	(0.590)	(0.300)	(0.300)
Covariates	Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year		Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year Time Trend	Publication Year Publication's Citations Publication's # Authors Scholar's Rank Scholar's Gender Accusations Year Time Trend
Observations	294,026	294,026	294,026	294,026
Random Effect			Scholar	Field
RE Groups			172	18
RE Variance			222.00	223.57
R2	0.742	0.001	0.742	0.741
Adjusted R2	0.742	0.001	0.745	0.742
Residual Std. Error	15.000 (df = 293,993)	29.400 (df = 294,018)	14.89	14.95

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 3. Regression table for Model 2 – Regression robustness checks[11].

[11] Location of models' descriptions – No time trend (1): Fig. 6; No covariates (2): Fig. 9; Scholas RE (3): Fig. 12; Field RE (4): Fig. 15.

Model 3: 2(Retraction Status: No vs. Yes) x 2(Timing: Pre- vs. Post-Accusation)

	Average Yearly Citations Per Publication			
	No time trend (1)	No covariates (2)	Scholars RE (3)	Field RE (4)
Post-accusation	-0.798***	-0.401	-1.814***	-1.799***
	(0.134)	(0.263)	(0.218)	(0.218)
Retraction	1.080**	0.388	0.965*	0.964**
	(0.492)	(0.977)	(0.494)	(0.491)
Post-accusation:Retraction	-1.573**	-2.149	-1.287*	-1.403*
	(0.729)	(1.465)	(0.730)	(0.729)
Covariates	Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year		Publication Year Publication's Citations Publication's # Authors Scholar's Field Scholar's Rank Scholar's Gender Accusations Year Time Trend	Publication Year Publication's Citations Publication's # Authors Scholar's Rank Scholar's Gender Accusations Year Time Trend
Observations	28,587	28,587	28,587	28,587
Random Effect			Scholar	Field
RE Groups			172	18
RE Variance			103.06	103.12
R2	0.754	0.000	0.751	0.753
Adjusted R2	0.753	0.000	0.755	0.753
Residual Std. Error	10.200 (df = 28,569)	20.500 (df = 28,583)	10.15	10.16

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 4. Regression table for Model 3 – Regression robustness checks[12].

[12] Location of models' descriptions – No time trend (1): Fig. 7; No covariates (2): Fig. 10; Scholas RE (3): Fig. 13; Field RE (4): Fig. 16.

Not controlling for time trend

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, and the year the accusations became public. While the citation rates of the controls scholars significantly increased after their respective accused scholar's allegations became public (b = 0.81, t(293,997) = 11.99, p < .001), those of the accused scholars decreased substantially more after the accusations became public (b = -1.25, t(293,997) = -8.32, p < .001).

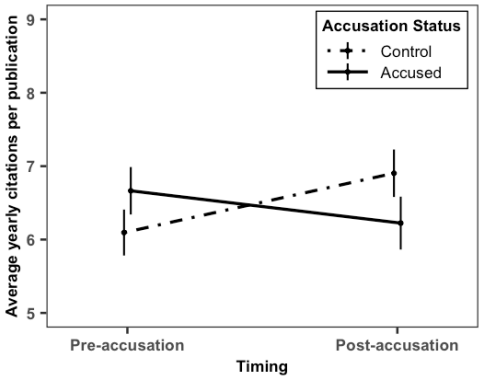


Fig. 5. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, not controlling for citations' trend.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend as fixed effects, and Scholar as random effect. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.69, t(293,993) = -7.96, p < .001) and those accused of scientific fraud (b = -0.80, t(293,993) = -3.77, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.89, t(293,993)= -2.97, p = .003) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

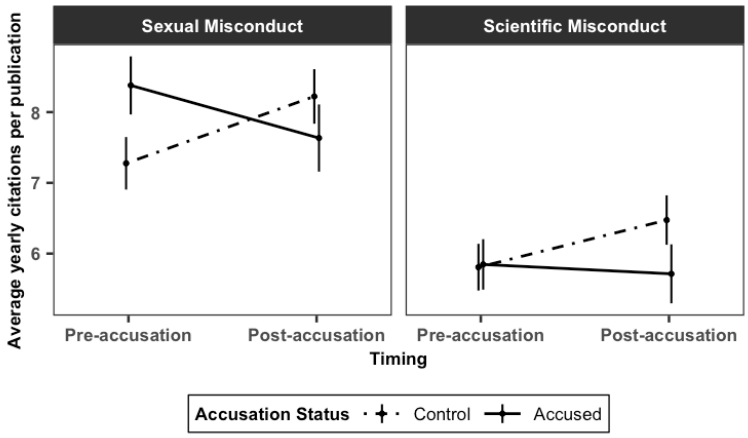


Fig. 6. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), not controlling for citations' trend.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, and the year the accusations became public. The citations of both retracted (b = -2.37, t(28,569) = -3.30, p < .001), and non-retracted publications (b = -0.80, t(28,569) = -5.94, p < .001) decreased after the accusations became public. The citations of the retracted publications decreased more than those of the non-retracted ones (b = -1.57, t(28,569) = -2.16, p = .031).

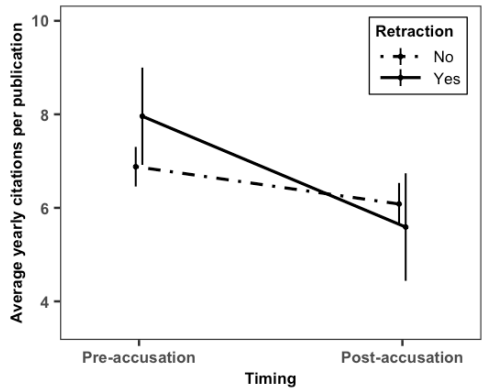


Fig. 7. Average yearly citations per publication by retraction status (scientific misconduct only), not controlling for citations' trend.

No covariates

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction. While the citation rates of the controls scholars significantly increased after their respective accused scholar's allegations became public (b = 1.11, t(294,022) = 8.56, p < .001), those of the accused scholars increased substantially less after the accusations became public (b = -1.53, t(294,022) = -5.19, p < .001).

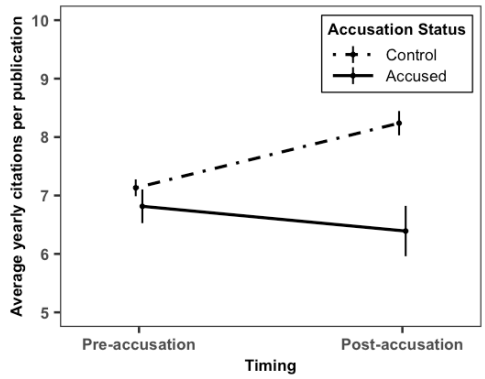


Fig. 8. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, not controlling for any covariate.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.28, t(294,018) = -3.06, p = .002) and those accused of scientific fraud (b = -1.85, t(294,018) = -4.43, p < .001) decreased, relative to their controls. The three-way interaction is not significant (b = 0.57, t(294,018)= 0.96, p = .335).

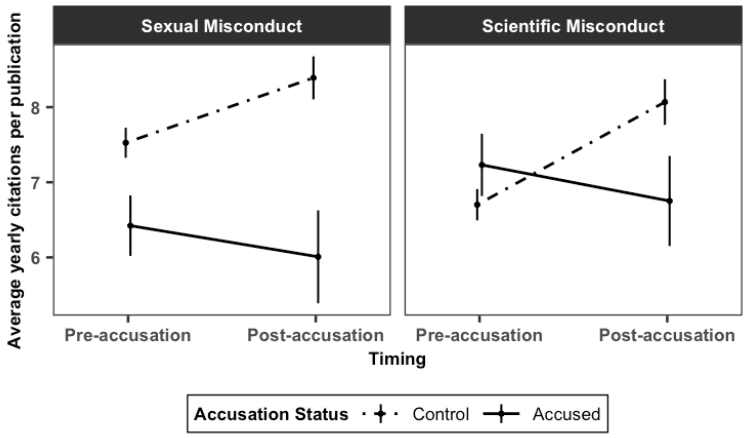


Fig. 9. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), not controlling for any covariate.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction. The citations of the retracted publications marginally decreased after the accusations became public (b = -2.55, t(28,583) = -1.77, p = .077), and those of the non-retracted publications remained constant (b = -0.40, t(28,583) = -1.52, p = .13). The interaction did not reach significance (b = -2.15, t(28,583) = -1.47, p = .14).

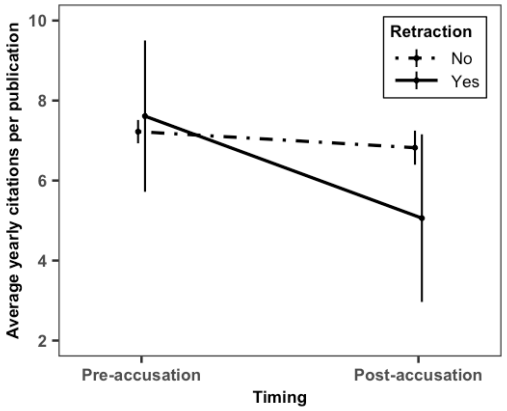


Fig. 10. Average yearly citations per publication by retraction status (scientific misconduct only), not controlling for any covariate.

With random effects for Scholar

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend as fixed effects, and Scholar as random effect. While the citation rates of scholars not accused of any misconduct remained unchanged before and after their respective accused scholar's allegations became public (b = 0.07, t(293,936.99) = 0.71, p = .477), those of the accused scholars decreased substantially more after the accusations became public (b = -1.17, t(293,778.06) = -7.80, p < .001).

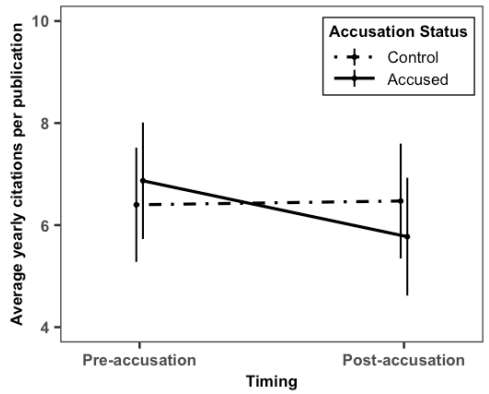


Fig. 11. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, with random effects for Scholar.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend as fixed effects, and Scholar as random effect. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.62, t(293,859.63) = -7.65, p < .001) and those accused of scientific fraud (b = -0.71, t(293,656.45) = -3.34, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.91, t(293,769.56)= -3.05, p = .002) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

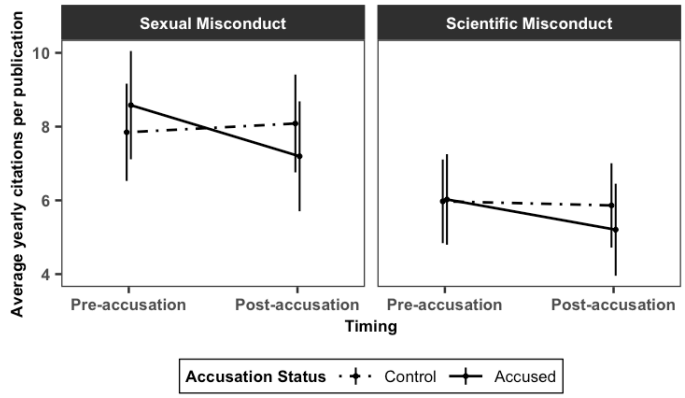


Fig. 12. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), with random effects for Scholar.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend as fixed effects, and Scholar as random effect. The citations of both retracted (b = -3.10, t(28,554.45) = -4.24, p < .001) and non-retracted publications (b = -1.81, t(28,567.56) = -8.31, p = .051) decreased. The citations of the retracted publications decreased marginally more than those of the non-retracted ones (b = -1.29, t(28,543.10) = -1.76, p = .078).

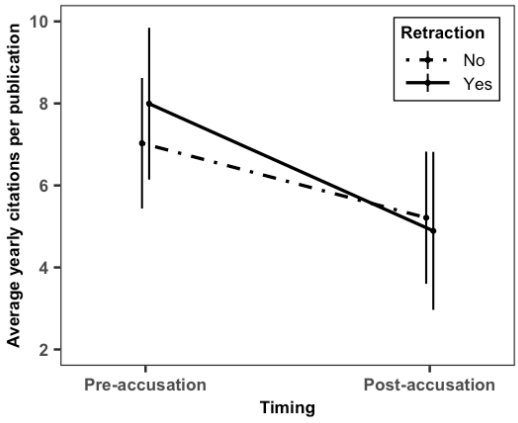


Fig. 13. Average yearly citations per publication by retraction status (scientific misconduct only), with random effects for Scholar.

With random effects for Field

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, rank, gender, the year the accusations became public, and time trend as fixed effects, and the scholar's discipline as random effect. While citations of controls remained unchanged (b = 0.08, t(293,998.65) = 0.73, p = .468), those of the accused scholars decreased substantially more after the accusations became public (b = -1.25, t(294,008.80) = -8.31, p < .001).

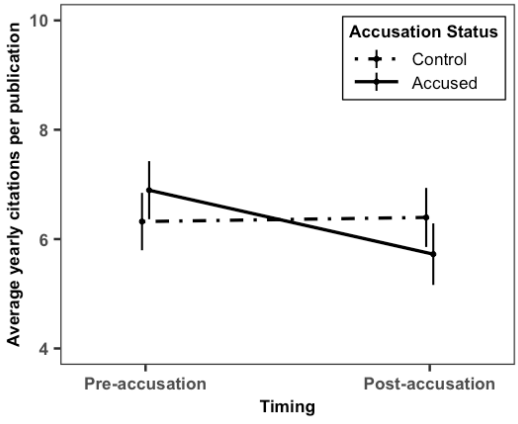


Fig. 14. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, with random effects for Field.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, rank, gender, the year the accusations became public, and time trend as fixed effects, and the scholar's discipline as random effect. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.70, t(293,997.28) = -8.02, p < .001) and those accused of scientific fraud (b = -0.78, t(294,005.96) = -3.68, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.92, t(294,003.00)= -3.06, p = .002) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

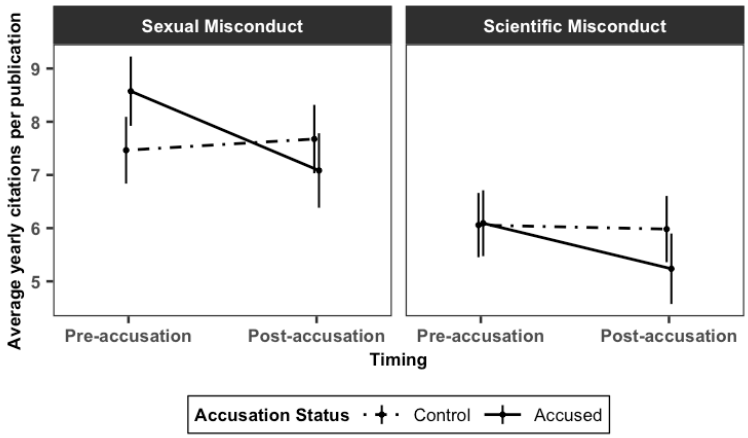


Fig. 15. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), with random effects for Field.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, rank, gender, the year the accusations became public, and time trend as fixed effects, and the scholar's discipline as random effect. The citations of both retracted (b = -3.20, t(28,567.11) = -4.38, p < .001) and non-retracted publications (b = -1.80, t(28,568.98) = -8.24, p < .001) decreased. The citations of the retracted publications decreased marginally more than those of the non-retracted ones (b = -1.40, t(28,565.61) = -1.93, p = .054).

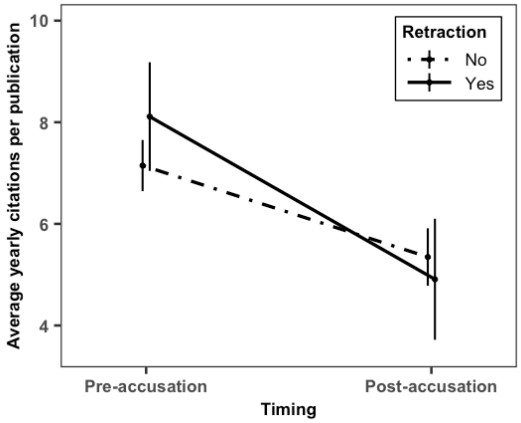


Fig. 16. Average yearly citations per publication by retraction status (scientific misconduct only), with random effects for Field.

SAMPLE ROBUSTNESS CHECKS

In this section we test our findings from Model 1, Model 2, and Model 3 across different sample's specifications. Specifically, we ran each of our models on five different samples:

- Including all 150 controls we initially selected without knowing 8 of them had at least one retracted publication (i.e., 150 controls)
- Excluding all publications that were published after the year the accused scholars were accused of misconduct (i.e., No publ. after accusation)
- Excluding all retracted publications (i.e., No retractions)
- Excluding the only scholar (and their controls) accused of scientific misconduct without any retracted publication—as they were accused of self-plagiarism (i.e., No self-plagiarism)

Regression Tables[13]

Model 1: 2(Accusation Status: Control vs. Accused) x 2(Timing: Pre- vs. Post-Accusation)

	Average Yearly Citations Per Publication			
	150 controls (1)	No publ. after accusations (2)	No retractions (3)	No self-plagiarism (4)
Post-accusation	0.144	0.390***	0.075	0.081
	(0.112)	(0.106)	(0.104)	(0.106)
Accused	0.673***	0.594***	0.562***	0.612***
	(0.094)	(0.087)	(0.087)	(0.088)
Post-accusation:Accused	-1.400***	-1.394***	-1.233***	-1.268***
	(0.164)	(0.153)	(0.151)	(0.153)
Observations	308,285	285,251	293,211	288,659
R2	0.752	0.748	0.614	0.742
Adjusted R2	0.752	0.748	0.614	0.742
Residual Std. Error	16.500 (df = 308,255)	14.900 (df = 285,221)	15.000 (df = 293,181)	15.100 (df = 288,630)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 17. Regression table for Model 1 – Sample robustness checks[14].

[13] All regressions control for the same covariates as the main analysis (i.e., controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend).

[14] Location of models' descriptions – 150 controls (1): Fig. 20; No publ. after accusations (2): Fig. 23; No retractions (3): Fig. 26; No self-plagiarism (4): Fig. 28.

Model 2: 2(Accusation Status: Control vs. Accused) x 2(Misconduct Type: Scientific vs. Sexual) x 2(Timing: Pre- vs. Post-Accusation)

	Average Yearly Citations Per Publication			
	150 controls (1)	No publ. after accusations (2)	No retractions (3)	No self-plagiarism (4)
Post-accusation	0.078	0.234*	-0.073	-0.056
	(0.133)	(0.128)	(0.125)	(0.128)
Accused	0.238*	0.067	-0.001	0.126
	(0.139)	(0.129)	(0.130)	(0.132)
Sexual	1.900***	1.472***	1.460***	1.494***
	(0.158)	(0.151)	(0.149)	(0.150)
Post-accusation:Accused	-1.032***	-0.918***	-0.759***	-0.834***
	(0.232)	(0.216)	(0.215)	(0.218)
Post-accusation:Sexual	0.131	0.294**	0.283**	0.256*
	(0.141)	(0.137)	(0.132)	(0.134)
Accused:Sexual	0.909***	1.054***	1.108***	0.970***
	(0.192)	(0.176)	(0.177)	(0.179)
Post-accusation:Accused:Sexual	-0.745**	-0.936***	-0.941***	-0.852***
	(0.328)	(0.306)	(0.302)	(0.305)
Observations	308,285	285,251	293,211	288,659
R2	0.752	0.748	0.742	0.742
Adjusted R2	0.752	0.748	0.742	0.742
Residual Std. Error	16.500 (df = 308,251)	14.900 (df = 285,217)	15.000 (df = 293,177)	15.100 (df = 288,626)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 18. Regression table for Model 2 – Sample robustness checks[11].

[11] Location of models' descriptions – 150 controls (1): Fig. 21; No publ. after accusations (2): Fig. 24; No retractions (3): Fig. 27; No self-plagiarism (4): Fig. 29.

Model 3: 2(Retraction Status: No vs. Yes) x 2(Timing: Pre- vs. Post-Accusation)

	Average Yearly Citations Per Publication			
	150 controls (1)	No publ. after accusations (2)	No retractions (3)	No self-plagiarism (4)
Post-accusation	-1.799***	-1.622***		-1.861***
	(0.218)	(0.221)		(0.224)
Retraction	0.961*	0.938*		0.892*
	(0.492)	(0.495)		(0.497)
Post-accusation:Retraction	-1.405*	-1.566**		-1.356*
	(0.729)	(0.732)		(0.735)
Observations	28,587	28,177		27,761
R2	0.754	0.754		0.756
Adjusted R2	0.754	0.754		0.755
Residual Std. Error	10.200 (df = 28,568)	10.200 (df = 28,158)		10.200 (df = 27,743)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 19. Regression table for Model 3 – Sample robustness checks[12].

[12] Location of models' descriptions – 150 controls (1): Fig. 22; No publ. after accusations (2): Fig. 25; No self-plagiarism (4): Fig. 30.

Including all 150 controls

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.14, t(308,255) = 1.29, p = .197), those of the accused scholars decreased substantially more after the accusations became public (b = -1.40, t(308,255) = -8.52, p < .001).

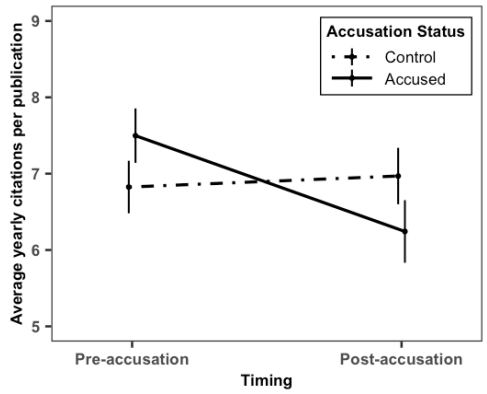


Fig. 20. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, including all 150 controls.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.78, t(308,251) = -7.63, p < .001) and those accused of scientific fraud (b = -1.03, t(308,251) = -4.45, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.74, t(308,251)= -2.27, p = .023) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

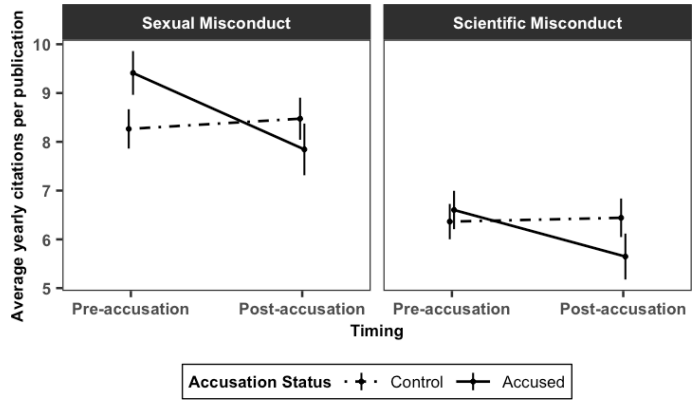


Fig. 21. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), including all 150 controls.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -3.20, t(28,568) = -4.38, p < .001) and non-retracted publications (b = -1.80, t(28,568) = -8.24, p < .001) decreased. The citations of the retracted publications decreased marginally more than those of the non-retracted ones (b = -1.40, t(28,568) = -1.93, p = .054).

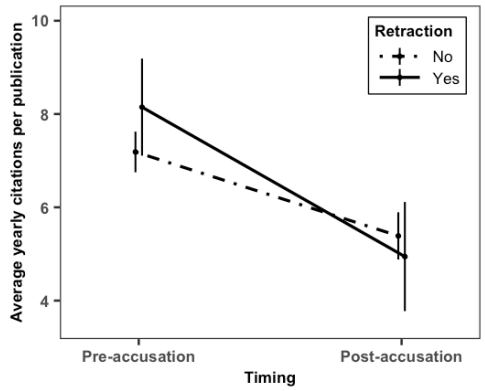


Fig. 22. Average yearly citations per publication by retraction status (scientific misconduct only), including all 150 controls.

Excluding publications published after the accusations became public

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars significantly increased after their respective accused scholar's allegations became public (b = 0.39, t(285,221) = 3.69, p < .001), those of the accused scholars decreased substantially more after the accusations became public (b = -1.39, t(285,221) = -9.13, p < .001).

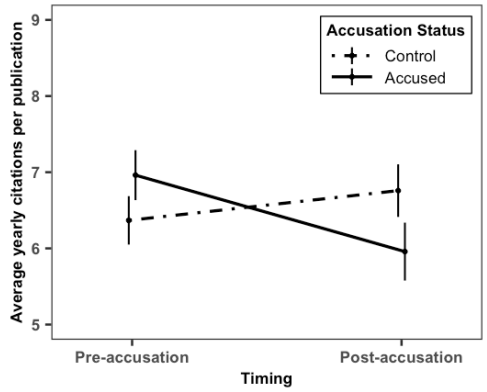


Fig. 23. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, excluding publications published after the accusations became public.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.85, t(285,217) = -8.58, p < .001) and those accused of scientific fraud (b = -0.92, t(285,217) = -4.25, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.94, t(285,217)= -3.06, p = .002) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

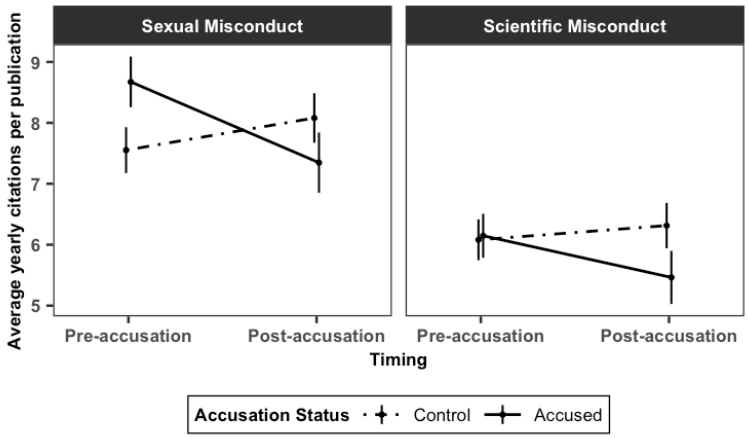


Fig. 24. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), excluding publications published after the accusations became public.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -3.19, t(28,158) = -4.34, p < .001) and non-retracted publications (b = -1.62, t(28,158) = -7.35, p < .001) decreased. The citations of the retracted publications decreased more than those of the non-retracted ones (b = -1.57, t(28,158) = -2.14, p = .033).

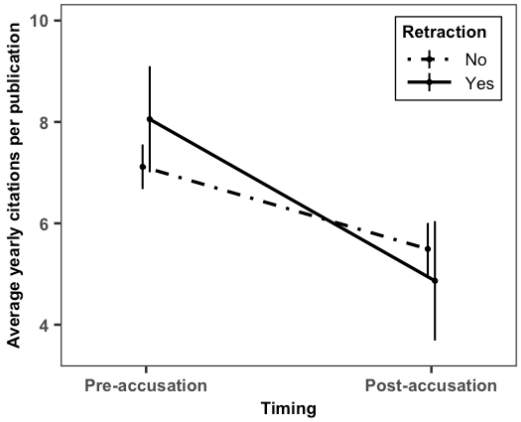


Fig. 25. Average yearly citations per publication by retraction status (scientific misconduct only), excluding publications published after the accusations became public.

Excluding retracted publications

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.08, t(293,181) = 0.72, p = .471), those of the accused scholars decreased substantially more after the accusations became public (b = -1.23, t(293,181) = -8.16, p < .001).

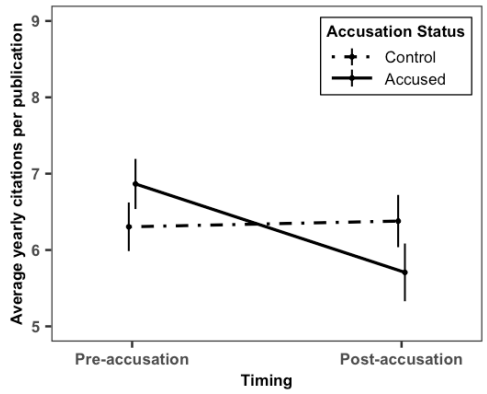


Fig. 26. Change in average yearly citations per publication by accusation status from before to after the news of the misconduct accusations became public, excluding retracted publications.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.70, t(293,177) = -8.01, p < .001) and those accused of scientific fraud (b = -0.76, t(293,177) = -3.52, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.94, t(293,177) = -3.11, p = .002) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

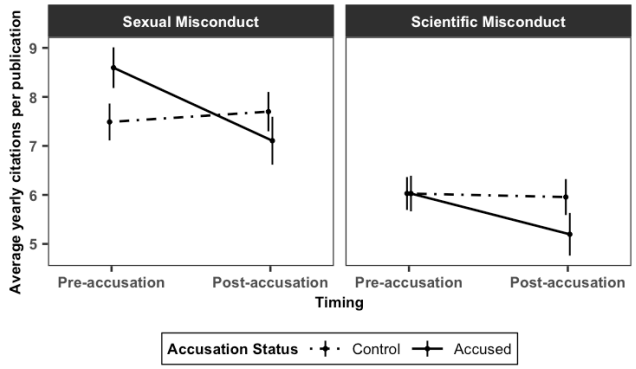


Fig. 27. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), excluding retracted publications.

- Model 3 -

In this robustness test we exclude all the retracted papers, consequently, we cannot run this analysis as it contrasts retracted vs. not retracted papers.

Excluding the only scholar in our sample accused of scientific misconduct who had no retractions—since they were accused of self-plagiarism

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.08, t(288,630) = 0.76, p = .446), those of the accused scholars decreased substantially more after the accusations became public (b = -1.27, t(288,630) = -8.31, p < .001).

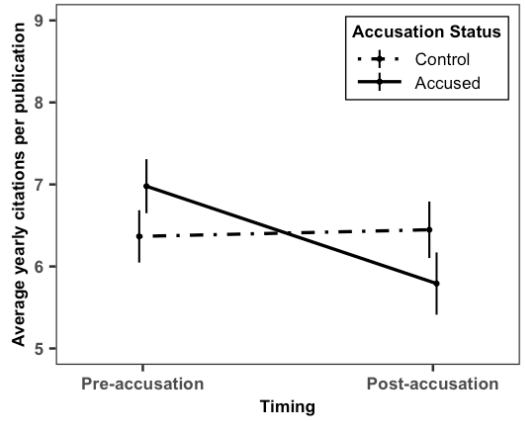


Fig. 28. Change in citations by accusation status from before to after the misconduct accusations became public, excluding the only scholar accused of scientific misconduct without retractions.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.69, t(288,626) = -7.88, p < .001) and those accused of scientific fraud (b = -0.83, t(288,626) = -3.83, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.85, t(288,626) = -2.79, p = .005) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

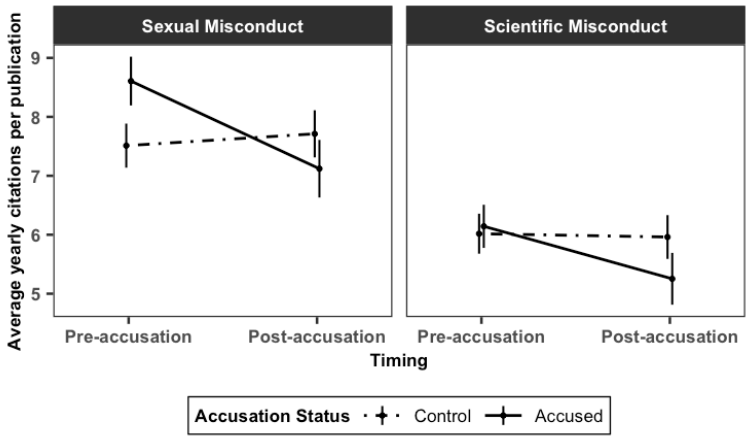


Fig. 29. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), excluding the only scholar in our sample accused of scientific misconduct who had no retractions.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -3.22, t(27,743) = -4.36, p < .001) and non-retracted publications (b = -1.86, t(27,743) = -8.32, p < .001) decreased. The citations of the retracted publications decreased marginally more than those of the non-retracted ones (b = -1.36, t(27,743) = -1.84, p = .065).

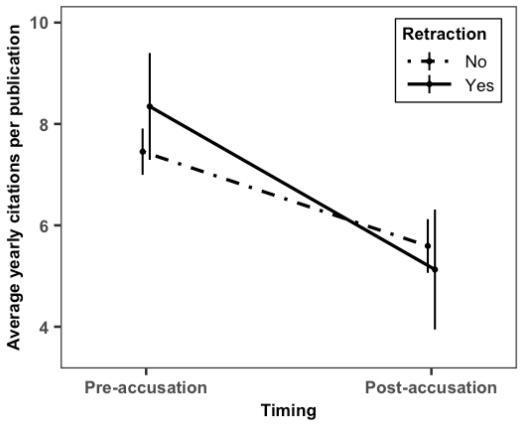


Fig. 30. Average yearly citations per publication by retraction status (scientific misconduct only), excluding the only scholar in our sample accused of scientific misconduct who had no retractions.

TIMING ROBUSTNESS CHECKS

In this section we test our findings from Model 1, Model 2, and Model 3 including different numbers of years in the 'Pre-accusation' period.
Specifically, we ran each of our models on four different time specifications:

- Including 5 years before the year the accusations became public, and 3 years after (i.e., 5 before – 3 after)
- Including 8 years before the year the accusations became public, and 3 years after (i.e., 8 before – 3 after)
- Including 12 years before the year the accusations became public, and 3 years after (i.e., 12 before – 3 after)
- Including 15 years before the year the accusations became public, and 3 years after (i.e., 15 before – 3 after)

Regression Tables[13]

MODEL 1: 2(Accusation Status: Control vs. Accused) x 2(Timing: Pre- vs. Post-accusation)

	Average Yearly Citations Per Publication			
	5 before - 3 after (1)	8 before - 3 after (2)	12 before - 3 after (3)	15 before - 3 after (4)
Post-accusation	0.167	-0.004	0.139	0.250***
	(0.153)	(0.116)	(0.097)	(0.091)
Accused	0.789***	0.664***	0.565***	0.514***
	(0.111)	(0.092)	(0.083)	(0.080)
Post-accusation:Accused	-1.427***	-1.331***	-1.241***	-1.192***
	(0.165)	(0.153)	(0.149)	(0.150)
Observations	213,258	265,814	317,548	345,637
R2	0.771	0.755	0.725	0.698
Adjusted R2	0.771	0.755	0.725	0.698
Residual Std. Error	15.000 (df = 213,228)	14.900 (df = 265,784)	15.100 (df = 317,518)	15.400 (df = 345,607)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 31. Regression table for Model 1 – Timing robustness checks[14].

[13] All regressions control for the same covariates as the main analysis (i.e., controlling for publication year, total citations per paper (as of June 2021), number of authors, scholar's discipline, rank (as of June 2021), gender, the year the accusations became public, and time trend).
[14] Location of models' descriptions – 5 before – 3 after (1): Fig. 34; 8 before – 3 after (2): Fig. 37; 12 before – 3 after (3): Fig. 40; 15 before – 3 after (4): Fig. 43.

MODEL 2: 2(Accusation Status: Control vs. Accused) x 2(Misconduct Type: Scientific vs. Sexual) x 2(Timing: Pre- vs. Post-accusation)

	Average Yearly Citations Per Publication			
	5 before - 3 after (1)	8 before - 3 after (2)	12 before - 3 after (3)	15 before - 3 after (4)
Post-accusation	0.121	-0.115	-0.056	-0.013
	(0.171)	(0.136)	(0.119)	(0.115)
Accused	0.599***	0.273	-0.054	-0.169
	(0.163)	(0.137)	(0.124)	(0.120)
Sexual	1.297***	1.409***	1.440***	1.336***
	(0.177)	(0.156)	(0.145)	(0.142)
Post-accusation:Accused	-1.313***	-1.004***	-0.699***	-0.585***
	(0.233)	(0.216)	(0.211)	(0.212)
Post-accusation:Sexual	0.091	0.212	0.368***	0.496***
	(0.146)	(0.135)	(0.131)	(0.131)
Accused:Sexual	0.419*	0.804***	1.223***	1.335***
	(0.225)	(0.188)	(0.169)	(0.164)
Post-accusation:Accused:Sexual	-0.235	-0.653**	-1.067***	-1.318***
	(0.330)	(0.305)	(0.298)	(0.251)
Observations	213,258	265,814	317,548	345,637
R2	0.772	0.755	0.725	0.698
Adjusted R2	0.772	0.755	0.725	0.698
Residual Std. Error	15.000 (df = 213,224)	14.900 (df = 265,780)	15.100 (df = 317,514)	15.400 (df = 345,603)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 32. Regression table for Model 2 – Timing robustness checks[17].

[17] Location of models' descriptions – 5 before – 3 after (1): Fig. 35; 8 before – 3 after (2): Fig. 38; 12 before – 3 after (3): Fig. 41; 15 before – 3 after (4): Fig. 44.

MODEL 3: 2(Retraction Status: No vs. Yes) x 2(Timing: Pre- vs. Post-accusation)

	Average Yearly Citations Per Publication			
	5 before - 3 after (1)	8 before - 3 after (2)	12 before - 3 after (3)	15 before - 3 after (4)
Post-accusation	-0.995***	-1.700***	-1.453***	-0.984***
	(0.286)	(0.231)	(0.202)	(0.186)
Retraction	0.385	1.102**	0.923*	0.973**
	(0.499)	(0.478)	(0.486)	(0.481)
Post-accusation:Retraction	-0.698	-1.501**	-1.391*	-1.432**
	(0.691)	(0.698)	(0.725)	(0.720)
Observations	21,244	26,149	30,522	32,635
R2	0.821	0.790	0.743	0.731
Adjusted R2	0.821	0.790	0.743	0.731
Residual Std. Error	9.070 (df = 21,225)	9.610 (df = 26,130)	10.200 (df = 30,503)	10.100 (df = 32,616)

Note: *p<0.1; **p<0.05; ***p<0.01

Fig. 33. Regression table for Model 3 – Timing robustness checks[18].

[18] Location of models' descriptions – 5 before – 3 after (1): Fig. 36; 8 before – 3 after (2): Fig. 39; 12 before – 3 after (3): Fig. 42; 15 before – 3 after (4): Fig. 45.

5 years before – 3 years after

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.17, t(213,228) = 1.09, p = .275), those of the accused scholars decreased substantially more after the accusations became public (b = -1.43, t(213,228) = -8.64, p < .001).

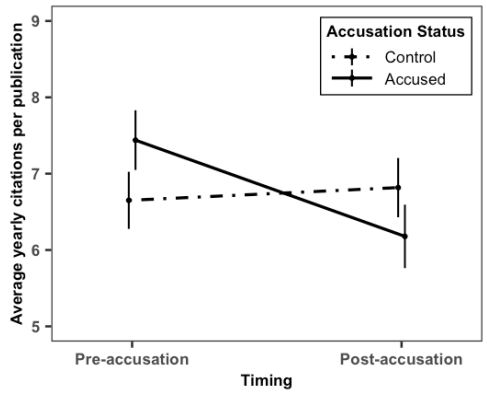


Fig. 34. Change in average yearly citations per publication by accusation status from before to after the accusations became public, including 5 years before and 3 years after the year the accusations broke.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.55, t(213,224) = -6.62, p < .001) and those accused of scientific fraud (b = -1.31, t(213,224) = -5.63, p < .001) decreased, relative to their controls. The three-way interaction does not reach significance (b = -0.24, t(213,224) = -0.71, p = .477).

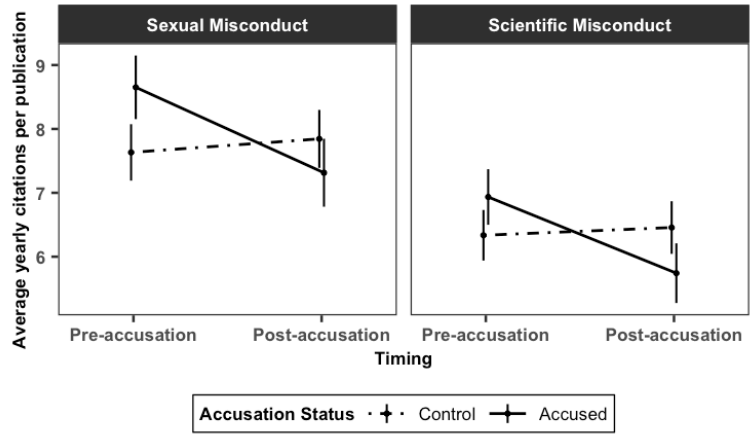


Fig. 35. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), including 5 years before and 3 years after the year the accusations became public.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -1.69, t(21,225) = -2.35, p = .019) and non-retracted publications (b = -0.99, t(21,225) = -3.48, p < .001) decreased. The citations of the retracted publications decreased as much as those of the non-retracted ones (b = -0.70, t(21,225) = -1.01, p = .312).

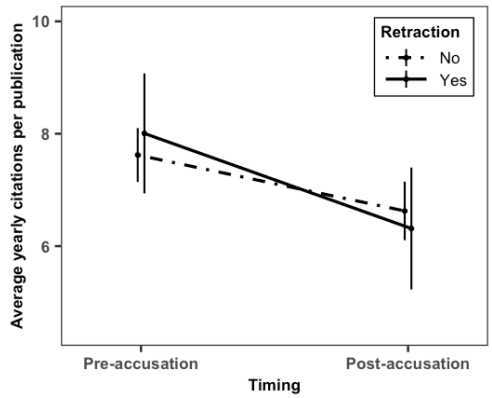


Fig. 36. Average yearly citations per publication by retraction status (scientific misconduct only), including 5 years before and 3 years after the year the accusations became public.

8 years before – 3 years after

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = -0.004, t(265,784) = -0.04, p = .972), those of the accused scholars decreased substantially more after the accusations became public (b = -1.33, t(265,784) = -8.71, p < .001).

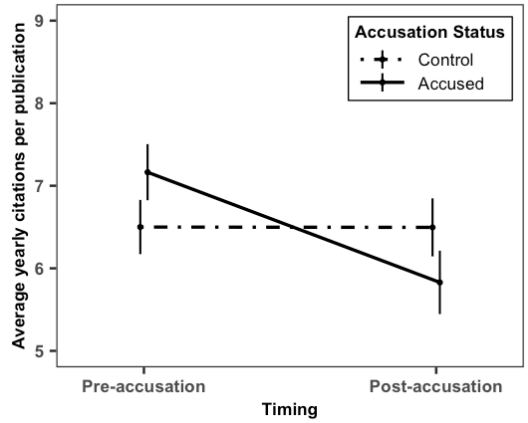


Fig. 37. Change in average yearly citations per publication by accusation status from before to after the misconduct accusations became public, including 8 years before and 3 years after the year the accusations became public.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.66, t(265,780) = -7.66, p < .001) and those accused of scientific fraud (b = -1.00, t(265,780) = -4.65, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -0.65, t(265,780) = -2.14, p = .033) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

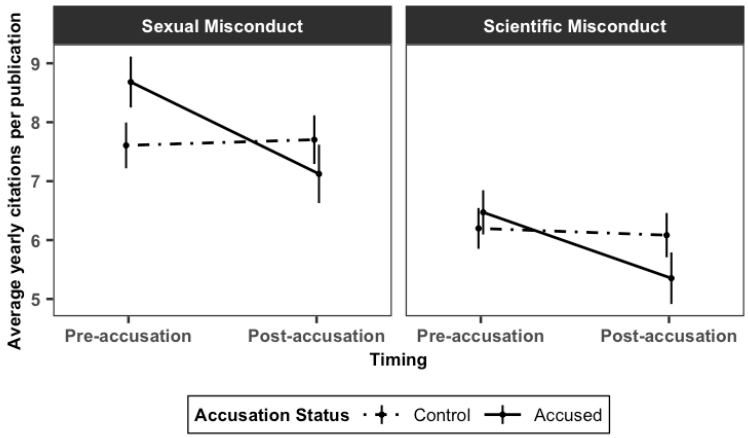


Fig. 38. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), including 8 years before and 3 years after the year the accusations became public.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -3.20, t(26,130) = -4.53, p < .001) and non-retracted publications (b = -1.70, t(26,130) = -7.35, p < .001) decreased. The citations of the retracted publications decreased more than those of the non-retracted ones (b = -1.50, t(26,130) = -2.15, p = .032).

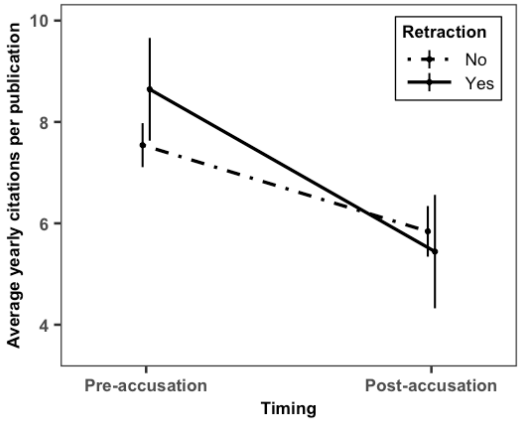


Fig. 39. Average yearly citations per publication by retraction status (scientific misconduct only), including 8 years before and 3 years after the year the accusations became public.

12 years before – 3 years after

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars remained unchanged before and after their respective accused scholar's allegations became public (b = 0.14, t(317,518) = 1.43, p = .153), those of the accused scholars decreased substantially more after the accusations became public (b = -1.24, t(317,518) = -8.32, p < .001).

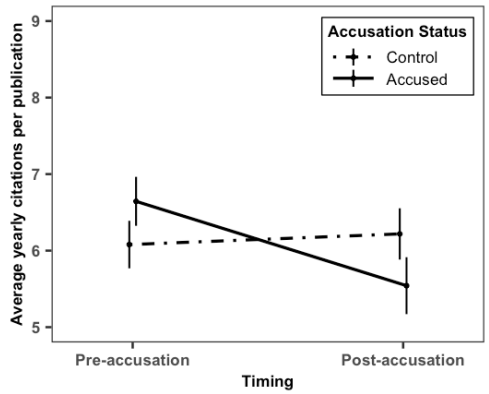


Fig. 40. Change in average yearly citations per publication by accusation status from before to after the misconduct accusations became public, including 12 years before and 3 years after the year the accusations became public.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis reveals that the citations of scholars accused of sexual misconduct (b = -1.77, t(317,514) = -8.37, p < .001) and those accused of scientific fraud (b = -0.70, t(317,514) = -3.31, p < .001) decreased, relative to their controls. Also, a significant three-way interaction (b = -1.07, t(317,514) = -3.57, p < .001) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

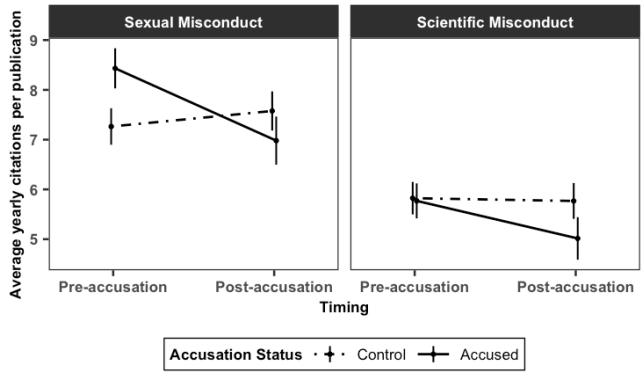


Fig. 41. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), including 12 years before and 3 years after the year the accusations became public.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -2.84, t(30,503) = -3.93, p < .001) and non-retracted publications (b = -1.45, t(30,503) = -7.20, p < .001) decreased. The citations of the retracted publications decreased marginally more than those of the non-retracted ones (b = -1.39, t(30,503) = -1.92, p = .055).

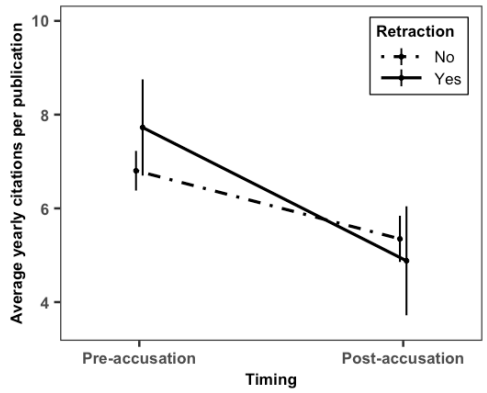


Fig. 42. Average yearly citations per publication by retraction status (scientific misconduct only), including 12 years before and 3 years after the year the accusations became public.

15 years before – 3 years after

- Model 1 -

We regress average yearly citations per publication on Accusation Status (Control vs. Accused), Time (Pre- vs. Post-Accusations), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. While the citation rates of the controls scholars significantly increased after their respective accused scholar's allegations became public (b = 0.25, t(345,607) = 2.73, p = .006), those of the accused scholars decreased substantially more after the accusations became public (b = -1.19, t(345,607) = -7.95, p < .001).

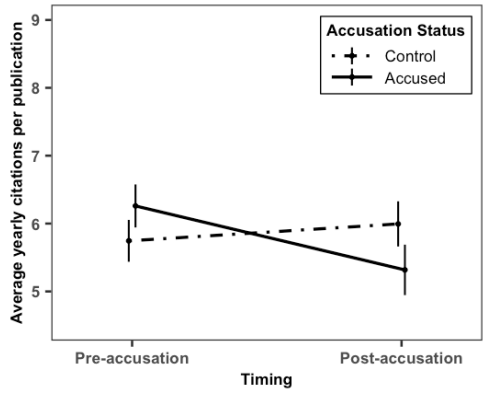


Fig. 43. Change in average yearly citations per publication by accusation status from before to after the accusations became public, including 15 years before and 3 years after the year the accusations broke.

- Model 2 -

We regress the average yearly citations per publication on Accusation Status (Control vs. Accused), Misconduct Type (Scientific vs. Sexual), Timing (Pre- vs. Post-Accusations), and their interactions, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. Our analysis revealed that the citations of scholars accused of sexual misconduct (b = -1.77, t(345,603) = -8.34, p < .001) and those accused of scientific fraud (b = -0.59, t(345,603) = -2.76, p = .006) decreased, relative to their controls. Also, a significant three-way interaction (b = -1.18, t(345,603) = -3.94, p < .001) reveals that the difference in citations between the scholars accused of sexual misconduct and their own controls was greater than that between scholars accused of scientific fraud and their own controls.

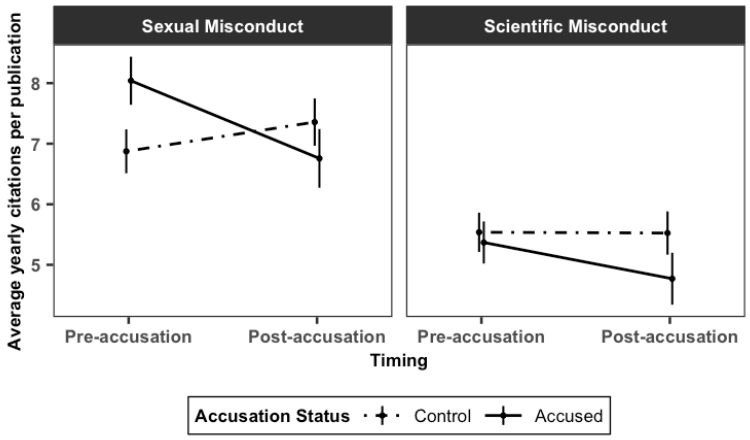


Fig. 44. Average yearly citations per publication by accusation status (accused vs. control), misconduct type (sexual vs. scientific), and timing (pre-accusation vs. post-accusation), including 15 years before and 3 years after the year the accusations became public.

- Model 3 -

We regress average yearly citations of researchers accused of scientific misconduct on Retraction Status (Yes vs. No), Timing (Pre- vs. Post-accusation), and their interaction, controlling for publication year, total citations per paper, number of authors, scholar's discipline, rank, gender, the year the accusations became public, and time trend. The citations of both retracted (b = -2.42, t(32,616) = -3.38, p < .001) and non-retracted publications (b = -0.98, t(32,616) = -5.30, p < .001) decreased. The citations of the retracted publications decreased more than those of the non-retracted ones (b = -1.43, t(32,616) = -1.99, p = .047).

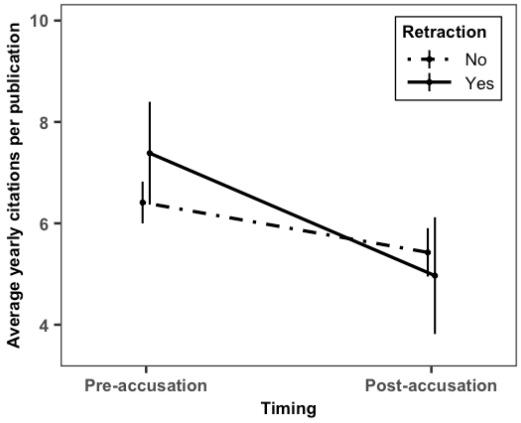
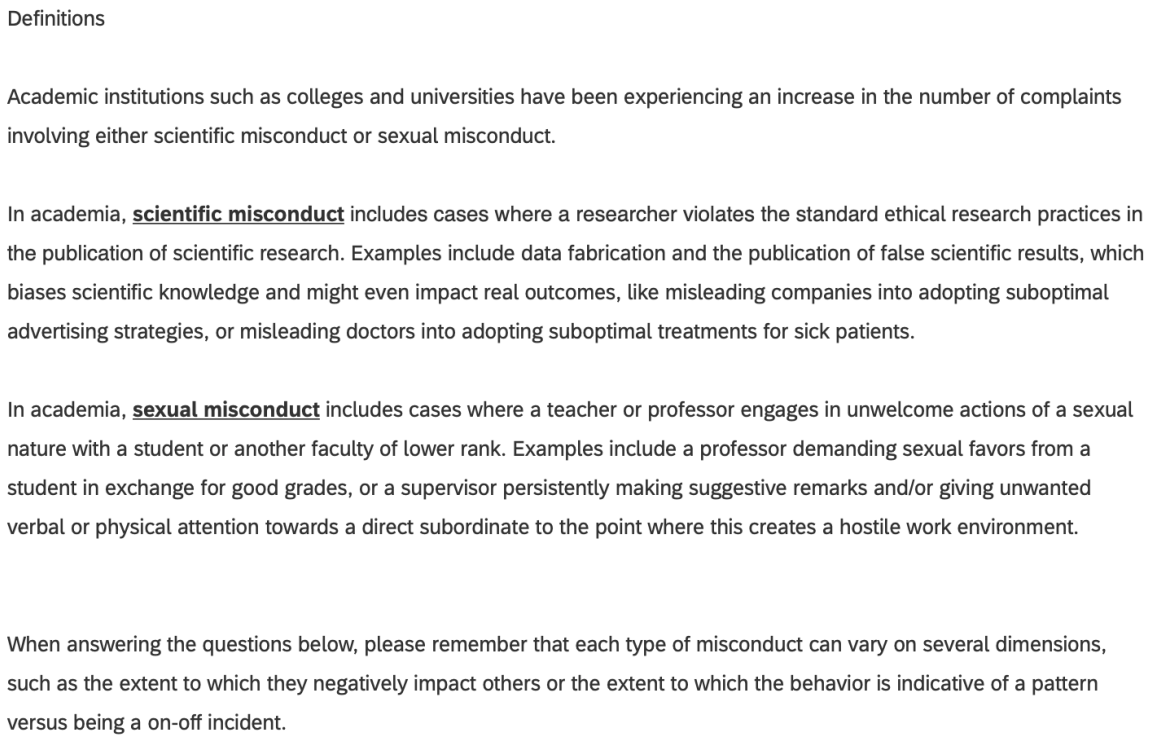


Fig. 45. Average yearly citations per publication by retraction status (scientific misconduct only), including 15 years before and 3 years after the year the accusations became public.

SURVEY[19]

Material

In the survey we first defined scientific and sexual misconduct in academia.


Definitions

Academic institutions such as colleges and universities have been experiencing an increase in the number of complaints involving either scientific misconduct or sexual misconduct.

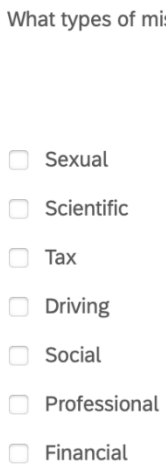
In academia, **scientific misconduct** includes cases where a researcher violates the standard ethical research practices in the publication of scientific research. Examples include data fabrication and the publication of false scientific results, which biases scientific knowledge and might even impact real outcomes, like misleading companies into adopting suboptimal advertising strategies, or misleading doctors into adopting suboptimal treatments for sick patients.

In academia, **sexual misconduct** includes cases where a teacher or professor engages in unwelcome actions of a sexual nature with a student or another faculty of lower rank. Examples include a professor demanding sexual favors from a student in exchange for good grades, or a supervisor persistently making suggestive remarks and/or giving unwanted verbal or physical attention towards a direct subordinate to the point where this creates a hostile work environment.

When answering the questions below, please remember that each type of misconduct can vary on several dimensions, such as the extent to which they negatively impact others or the extent to which the behavior is indicative of a pattern versus being a on-off incident.

Fig. 46. Definition of scientific and sexual misconduct in academia given in the survey.

On a separate page, we asked participants to indicate what types of misconduct they just read about


What types of misconduct did we just describe? (pick two)

- ☐ Sexual
- ☐ Scientific
- ☐ Tax
- ☐ Driving
- ☐ Social
- ☐ Professional
- ☐ Financial

Fig. 47. Attention check question in the survey.

[19] Data, Material (qsf file), and R code for the analysis can be found on https://osf.io/ycazs/?view_only=bfb1080d756146ba89d21a7ed3daeacf .

Then, we asked participants to indicate which of the two types of misconduct is (a) more deserving of punishment, (b) more disgusting, and (c) worse than the other.

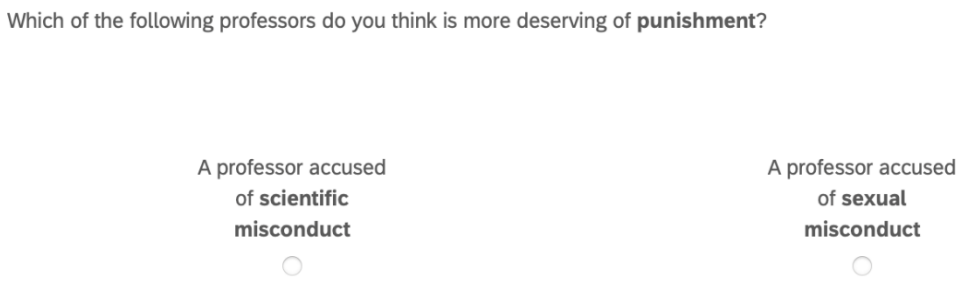


Fig. 48. Punishment question in the survey.

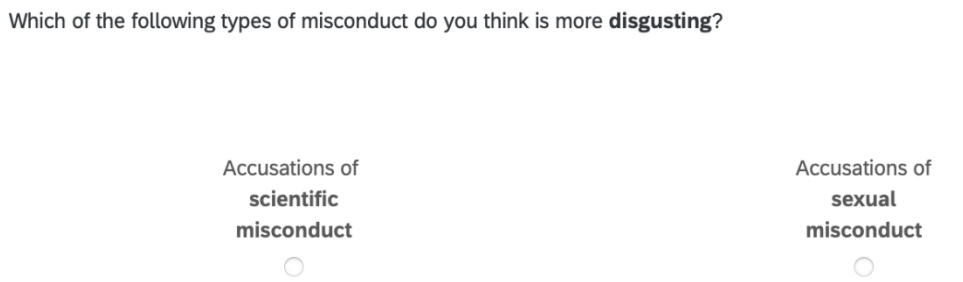


Fig. 49. Disgust question in the survey.

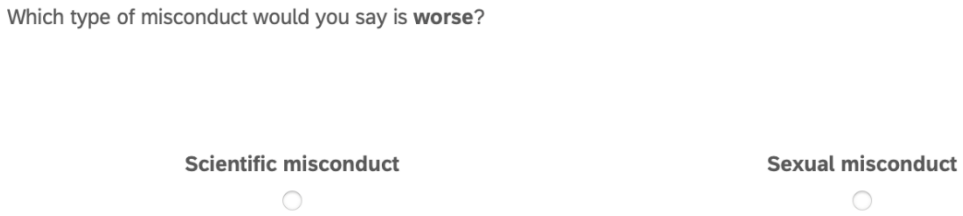


Fig. 50. Which misconduct type is worse question in the survey.

Note that, while the responses' order within each question was randomized, these three questions' order was *not* randomized. In the first and third questions (deserving punishment and being worse) sexual misconduct got almost the identical share of responses—suggesting people were consistent in indicating they would punish more the transgressor they thought was accused of the worse misconduct type. Also, roughly 13% of our participants indicated that even though they thought scientific misconduct was worse (question 3) and more deserving of punishment (question 1), sexual misconduct was still more disgusting (question 2), suggesting the disgust question did not solely influence participants' answers in the last question.

Results' Graphs

Analysis on overall sample (N=231)

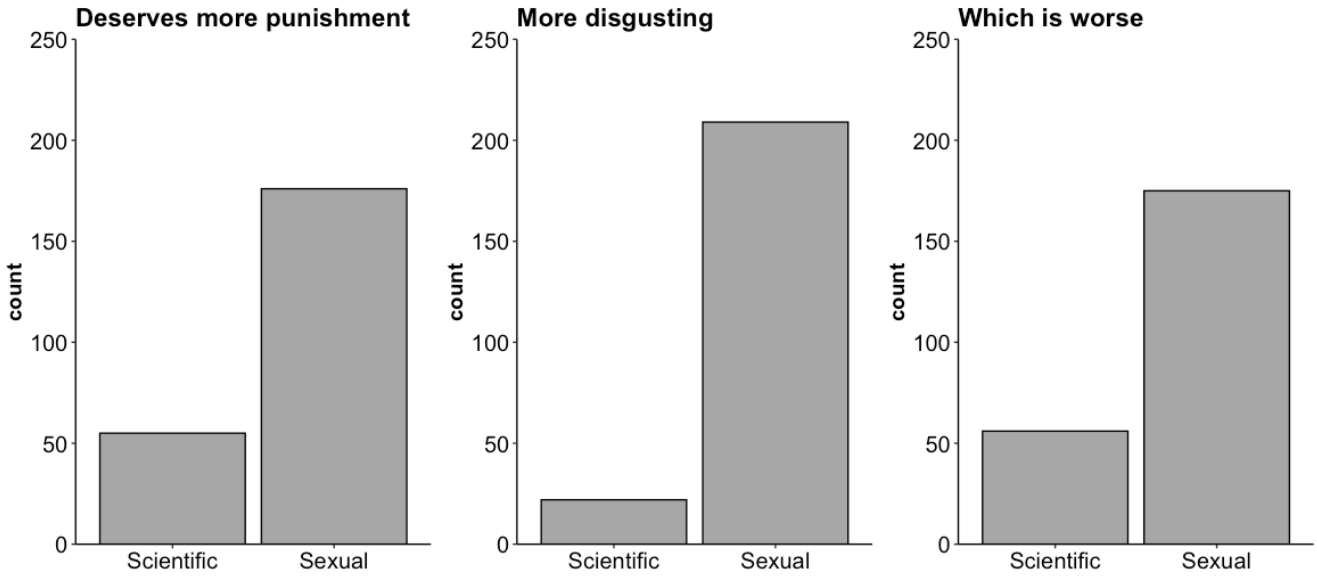


Fig. 51. Results on overall survey sample.

Analysis by Gender

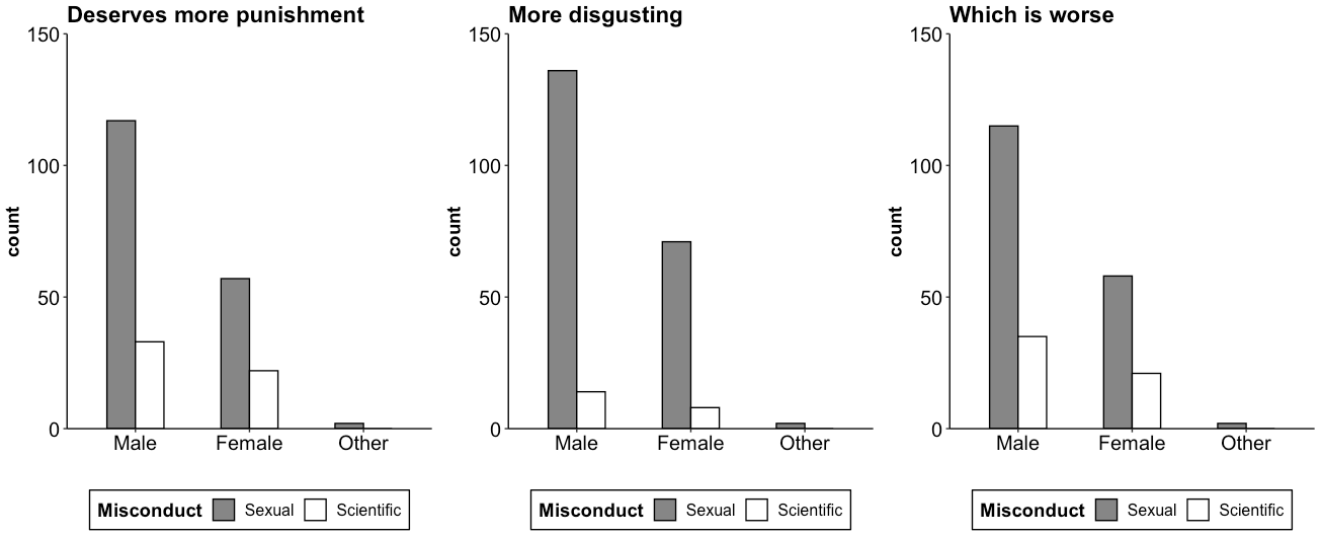


Fig. 52. Results by gender from the survey.

REFERENCES

1. E. McCarthy, "#MeToo Raised Awareness about Sexual Misconduct. Has It Curbed Bad Behavior?". *The Washington Post*. Retrieved April 2022 (2021).

2. R. Levy, M. Mattsson, "The Effects of Social Movements: Evidence from the #MeToo". Available at *SSRN: https://ssrn.com/abstract=3496903* (2019).

3. E. C. Tippett, "The Legal Implications of the MeToo Movement". *Minn. Law Rev.*, **103**, 229–302 (2018).

4. N. C. Cantalupo, W. C. Kidder, "A systematic look at a serial problem: Sexual harassment of students by university faculty". *2018 Utah L. Rev.*, 671–786 (2018).

5. A. E. Tenbrunsel, M. R. Rees, K. A. Diekmann, "Sexual harassment in academia: Ethical climates and bounded ethicality". *Annu. Rev. Psychol.*, **70**, 245–270 (2019).

6. M. Nussbaum, "Thoughts about sexual assault on college campuses". *Brookings*. Retrieved 5/22/2022. https://www.brookings.edu/research/thoughts-about-sexual-assault-on-college-campuses/

7. D. K. Chan, S. Y. Chow, C. B. Lam, S. F. Cheung, "Examining the job-related, psychological, and physical outcomes of workplace sexual harassment: A meta-analytic review". *Psychol. Women Q.*, **32**, 362–376 (2008).

8. E. Suran, "Title IX and Social Media: Going Beyond the Law". *Mich. J. Gender & L.*, **21**, 273–310 (2014).

9. K. Hyland, "Academic Attribution: Citation and the Construction of Disciplinary Knowledge". *Applied Linguistics*, **20**, 341–367 (1999).

10. C. Catalini, N. Lacetera, A. Oettl, "The incidence and role of negative citations in science". *Proc. Natl. Acad. Sci. USA,* **112**, 13823–13826 (2015).

11. K. Siler, K. Lee, L. Bero, "Measuring the effectiveness of scientific gatekeeping". *Proc. Natl. Acad. Sci. USA,* **112**, 360–365 (2015).

12. R. K. Merton (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press, Chicago, IL (1973).

13. R. K. Merton, "The Matthew Effect in Science II: Cumulative Advantage and the Symbolism of Intellectual Property". *Isis*, **79**, 606–623 (1988).

14. "What Does it Mean to Cite?". *MIT Academic Integrity.* Retrieved 5/9/2022. https://integrity.mit.edu/handbook/citing-your-sources/avoiding-plagiarism-cite-your-source

15. L. Bornmann, H. D. Daniel, “What do Citation Counts Measure? A Review of Studies on Citing Behavior”. *J. Doc*., **64**, 45–80 (2008).

16. K. O. May, N. C. Janke, “Abuses of Citation Indexing”. *Science*, **156**, 890–892 (1967).

17. S. F. Lu, G. Z. Jin, B. Uzzi, B. Jones, “The Retraction Penalty: Evidence from the Web of Science”. *Sci. Rep*., **3**, 1–5 (2013).

18. N. H. Steneck, “Introduction to the Responsible Conduct of Research”. *Washington, DC: US Government Printing Office* (2007).

19. C. Piller, “Blots on a Field? A Neuroscience Image Sleuth Finds Signs of Fabrication in Scores of Alzheimer’s Articles, Threatening a Reigning Theory of the Disease”. *Science*, **377**, 358–363 (2022).

20. J. Z. Berman, D. A. Small, “Discipline and Desire: On the Relative Importance of Willpower and Purity in Signaling Virtue”. J. Exp. Soc. Psychol., 76, 220-230 (2018).

21. D. Komić, S. L. Marušić, A. Marušić, “Research Integrity and Research Ethics in Professional Codes of Ethics: Survey of Terminology Used by Professional Organizations across Research Disciplines”. *PLoS ONE*, **10**, 1–13 (2015).

22. J. Libarkin, “Academic Sexual Misconduct Database”. Retrieved 11/22/2020, https://academic-sexual-misconduct-database.org

23. “Retraction Watch”. Retrieved 2/3/2022, https://retractionwatch.com

24. “List of Scientific Misconduct Incidents”. Retrieved 3/14/2021, https://en.wikipedia.org/wiki/List_of_scientific_misconduct_incidents

25. X. Bai, F. Zhang, I. Lee, “Predicting the Citations of Scholarly Paper”. *J. Informetr.*, **13**, 407–418 (2019).

26. J. Mingers, “Exploring the Dynamics of Journal Citations: Modelling with S-curves”. *J. Oper. Res. Soc.,* **59**, 1013–1025 (2008).

27. M. Serra-Garcia, U. Gneezy, “Nonreplicable Publications are Cited More than Replicable Ones”. *Sci. Adv.,* **7**, 1–7 (2021).

28. J. P. A. Ioannidis, “Why Science Is Not Necessarily Self-Correcting”. *Psychol. Sci.,* **7**, 645–654 (2012).

29. G. Nave, C. F. Camerer, M. McCullough, “Does Oxytocin Increase Trust in Humans? A Critical Review of Research”. *Perspect. Psychol. Sci.*, **10**, 772–789 (2015).

30. E. Vul, C. Harris, P. Winkielman, H. Pashler, "Puzzling High Correlations in fMRI Studies of Emotions, Personality, and Social Cognition". *Perspect. Psychol. Sci.*, **4**, 274–290 (2009).

31. C. F. Camerer, A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E. Wagenmakers, H. Wu, "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015". *Nat. Hum. Behav.*, **2**, 637–644 (2018).

32. C. F. Camerer, A. Dreber, E. Forsell, T. H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, H. Wu, "Evaluating replicability of laboratory experiments in economics". *Science*, **351**, 1433–1436 (2016).

33. J. P. Simmons, L. D. Nelson, U. Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". *Psychol. Sci.*, **22**, 1359–1366 (2011).

34. L. D. Nelson, J. P. Simmons, U. Simonsohn, "Psychology's Renaissance". *Annu. Rev. Psychol.*, **69**, 511–534 (2018).

35. D. A. Moore, "Preregister if you Want to". *Am. Psychol.*, **71**, 238–239 (2016).

36. S. Schnall, J. Haidt, G. L. Clore, A. H. Jordan, "Disgust as Embodied Moral Judgment". *Pers. Soc. Psychol. Bull.*, **34**, 1096–1109 (2008).

37. J. Haidt, "The Moral Emotions". In *Handbook of Affective Sciences*, eds. R. J. Davidson, K. R. Scherer, H. H. Goldsmith, Oxford: Oxford University Press, 852–870 (2003).

38. A. Marušić, E. Wager, A. Utrobicic, H. R. Rothstein, D. Sambunjak, "Interventions to Prevent Misconduct and Promote Integrity in Research and Publication". *Cochrane Database of Systematic Reviews*, **4**, 1–21 (2016).

39. P. Slovic, J. Flynn, H. Kunreuther (Ed.), *Risk, Media and Stigma: Understanding Public Challenges to Modern Science and Technology* Routledge, London, UK (2001).

40. "When It Comes to Sexual Harassment, Academia Is Fundamentally Broken". *Scientific American.* Retrieved 10/3/2022. https://blogs.scientificamerican.com/voices/when-it-comes-to-sexual-harassment-academia-is-fundamentally-broken/

Chapter 2.

NOT ALL ATTRIBUTIONS ARE SELF-SERVING: A PREFERENCE FOR AGENCY OVER NEGATIVE OUTCOMES

Giulia Maimone, Joachim Vosgerau, Ayelet Gneezy

Rady School of Management, University of California, San Diego, La Jolla, CA, 92093, USA

ABSTRACT

There are two streams of literature that address attributional preferences: self-determination and self-serving preferences. While these two theories make the same prediction for individuals' attributional preferences over positive outcomes, they make competing predictions for attributional preferences over negative outcomes. Self-determination maintains that people prefer to have agency over negative outcomes. Self-serving preferences, in contrast, stipulate that people prefer to concede agency over negative outcomes. In eight preregistered experiments (N = 3,946), we reconcile these seemingly inconsistent attributional preferences over negative outcomes. First, we test these competing predictions and find that—consistent with self-determination—people would rather "own" their negative outcomes than externally attribute them. Overplacement (people's belief that they perform better than others) and the impact bias (the belief that "owned" negative outcomes hurt less than when they are caused by oneself) cannot explain this preference. Instead, we find that reducing the saliency of agency moderates the preference for agency over negative outcomes. More interestingly, we find that sharing agency reverses attributional preferences: while people prefer assuming agency over negative outcomes when these are exclusively caused by a sole agent (either themselves or somebody else), they prefer attributing agency to others when negative outcomes are jointly caused by multiple agents (both themselves and somebody else).

During the fall of 2021, America was anxiously awaiting the verdict of a divisive high-profile court case—the case of Kyle Rittenhouse, a teenager who shot three men, killing two, during a protest against police brutality in Wisconsin. The defendant argued he fired in self-defense after the men attacked him. During the jury selection process, Bruce Schroeder, the judge in charge of the case, made a highly unusual decision, allowing Rittenhouse to draw numbers from a lottery tumbler to determine the alternate jurors. Explaining why he made this decision, Schroeder stated that, regardless of what the outcome would be, "people feel better when they have control."[15]

There are two theories on attributional preferences that can inform us about the correctness of Judge Schroeder's intuition. One stream of literature—*self-determination preferences*—argues that people generally prefer self-determination, autonomy (e.g., Deci & Ryan, 2000; 2013) and personal agency (Crisp & Barber, 1995; Gneezy, Imas, & Jaroszewicz, 2020) over their environment, irrespective of outcomes. According to theories on the perceived locus of causality (PLOC; Heider, 1958), however, people prefer agency over positive but not over negative outcomes. PLOC is the psychological place to which people assign causes of events, which can be internal or external. Under an internal PLOC, individuals attribute outcomes to their own choices and behaviors. Under an external PLOC, individuals attribute outcomes to external forces (e.g., situational factors, fate, or other individuals). *Self-serving preferences* occur as people are motivated to protect their self-esteem and maintain a positive self-image by attributing desirable outcomes internally and undesirable ones externally (cf., DeCharms, 1968; Heider, 1958; Kelley, 1967; Kelley, 1973; Larson, 1977; Miller & Ross,

[15] https://www.cnn.com/us/live-news/kyle-rittenhouse-trial-verdict-watch-11-17-21/h_04578c066a37ca5cacf3fa99b3ab1e2d; Judge Schroeder presumably referred to a subjective experience of illusionary control (Langer, 1975; Wortman, 1975).

1975). If Rittenhouse were acquitted (i.e., in case of positive outcome) both theories would predict that Judge Schroeder was right in assuming that Rittenhouse would prefer (illusory) agency over that outcome. If Rittenhouse were convicted (i.e., in case of negative outcome), however, the two theories would make opposing predictions. According to self-determination preferences, Judge Schroeder was correct in predicting Rittenhouse's preference. According to self-serving preferences, Judge Schroeder was *not* correct in predicting Rittenhouse's preference, as he would have preferred to not being responsible for such negative outcome.

In this paper, we investigate whether Judge Schroeder's intuition that people prefer owning an outcome—even if it turns out to be negative—was correct. Specifically, we test whether people prefer assuming agency over a negative outcome (i.e., an outcome that bears direct negative consequences for them), even when doing so reflects negatively on their skills and/or knowledge.

We find consistent support for such a preference for agency over negative outcomes, and test two alternative explanations besides a general preference for agency that may explain these preferences. The first, *overplacement* (Moore & Healy, 2008; Moore & Schatz, 2017), is a form of overconfidence that represents individuals' beliefs that they have more skills, knowledge, or luck than others. If overplacement were causing people to prefer agency over negative outcomes, they would do so because by acting themselves—regardless of how things turn out—they would minimize the probability of the negative outcome materializing. As an affective consequence, owning the negative outcome would minimize post-outcome counterfactual thinking (Frith, 2014). The second alternative explanation is the *impact bias*. When forecasting the affective impact of future events and experiences, individuals tend to overestimate their intensity (e.g., Gilbert et al., 1998; Gilbert et al., 2004a; Wilson & Gilbert, 2003), especially when considering

atypical instances. For example, people predict an accident to be more painful when it happened on an unusual, rather than on the usual, way to work (Gilbert et al., 2004b; Kahneman & Miller, 1986). If emotional responses to negative outcomes caused by others—which are arguably less common and therefore more atypical—result in more intense affective reactions than responses to negative outcomes caused by oneself, individuals may prefer agency over the negative outcome.

We find no evidence for these two alternative mechanisms. Instead, we hypothesize that—consistent with isolation effect in the "elimination by aspects" theory of choice (Tversky, 1972)—when making choices, people have a hierarchy of motives, and that the two most important motives for people expressing their attributional preferences are self-determination and self-enhancement concerns. We predict and find that self-determination—as a basic human need (Deci & Ryan, 2000)—is the first and most important aspect/attribute individuals consider when outcomes are caused exclusively by a sole agent. If people are asked to choose between attributing a negative outcome either to themselves *or* to somebody else, they prefer having agency and assuming responsibility for the outcome. However, when outcomes are caused jointly by multiple agents (i.e., themselves *and* someone else), people can satisfy their desire for self-determination and at the same time protect their self-esteem by assuming some agency over the outcome but attributing a larger share of it to others. Consistent with this account, we find that the preference for sole agency over negative outcomes is moderated when the decision-maker's attention is shifted away from the agent of that outcome, and that it reverses when agency is not exclusive but shared with others, resulting in self-serving attributions.

EMPIRICAL APPROACH

Across eight preregistered studies, we test a) whether a preference for agency over negative outcomes exists, b) two alternative psychological mechanisms besides a general preference for self-determination and autonomy that may underlie such preferences, and c) the boundary conditions under which attributions of negative outcomes are driven by a general desire for autonomy and self-determination or by a desire to protect one's self-esteem and maintain a positive self-image.

Attributional preferences can be measured by asking respondents to a) rate the amount of responsibility or effort exerted in producing an outcome that they ascribe to themselves and to others, b) choose among agency-outcome scenarios, and c) rate their satisfaction/dissatisfaction with agency-outcome scenarios. A disadvantage of the first method—rated responsibility/effort—is that it works only for outcomes caused by multiple agents, since in the case of a sole agent responsibility/effort can only be attributed to that one agent. A disadvantage of the second method—choosing among agency scenarios—is that it requires the direct comparison of agency scenarios from which respondents can choose. The third method—satisfaction/dissatisfaction ratings with agency scenarios—solves all these issues as it can be applied to both sole and multiple agency scenarios, as well as to scenarios that are evaluated jointly and separately. In our studies, we assess attributional preferences with choice shares and rated satisfaction/dissatisfaction levels for agency-outcome scenarios.

In study 1, we measure agency preferences for three affective-rich negative outcomes that people may encounter in their everyday lives and demonstrate a preference for personal agency over these outcomes. In study 2, we ask participants to choose among agency-scenarios and show that participants not only prefer receiving a negative outcome when it is caused by

themselves than when it is caused by someone else, but also when it is caused by chance—highlighting that such agency preference is not simply a distaste for blaming others, but more generally it is a preference for internal versus external attributions. Study 2 also rules out overplacement as the underlying mechanism for the observed preferences. In study 3, we test whether the impact bias can account for the findings of studies 1 and 2. We conclude that it cannot. According to our hypothesis that, motivated by a desire for self-determination, people exhibit a general preference for agency when outcomes are exclusively caused by a sole agent, in studies 4A, 4B and 5 we test and show that the preference for agency is attenuated when saliency is drawn away from the agent (e.g., towards the outcome). Specifically, in studies 4A and 4B we show that the preference for agency over negative outcomes is stronger in joint (where agent scenarios can be directly compared) than in separate evaluations (where agent scenarios cannot be directly compared). In study 5 we demonstrate an order effect whereby the preference for agency is diminished when respondents rate their satisfaction with a positive outcome before they rate their dissatisfaction with a negative outcome. Finally, in studies 6A and 6B we test whether a preference for self-determination occurs when outcomes are construed as being caused exclusively by a sole agent, but a preference for self-serving attributions occurs when outcomes are construed as being caused jointly by multiple agents.

All studies are preregistered; preregistrations[16], data, analyses, and experimental materials can be accessed here: https://researchbox.org/532&PEER_REVIEW_passcode=EZYJTY.

[16] In some preregistrations, we refer to a preference for agency over negative outcomes as "Die By My Own Hand" (DBMOH) preference.

STUDY 1

Study 1 was designed to test whether participants would prefer agency over 3 affective-rich negative outcomes that they may encounter in their everyday lives.

Method

Participants. Four hundred forty-nine CloudResearch approved Amazon Mechanical Turk workers from the US (47.2% female, 0% other, M_{Age} = 39.6 years, SD_{Age} = 11.4 years) completed the experiment in exchange for monetary compensation ($0.30 for 2 minutes expected completion time), one short of the 450 that we had preregistered. Participants who failed to correctly answer an attention-check question were not allowed to start the study.

Procedure. Study 1 employed a 3 (scenario: car accident vs. ski accident vs. law school dropout; between-subject) x 2 (agent: self vs. other; within-subject) design. Participants read one of three scenarios in which they were told to imagine experiencing a negative outcome, and were asked what they would prefer, experiencing the negative outcome because of a choice/action that they had carried out or because of a choice/action that someone else had carried out. Specifically, in the car scenario, participants imagined that they got into a car accident with their own car and indicated whether they preferred having gotten into the accident while they were driving or while a friend of theirs was driving. In the ski accident scenario, participants imagined having lost control of their skies and badly broken their right leg, and then indicated whether they preferred having lost control of their skis because they had accidentally slipped on an ice sheet or because another skier had accidentally hit them. In the law school scenario, participants were told that during her studies at law school, Emily had realized that she would never want to practice law and decided to drop out. Participants then indicated whether they thought Emily would prefer

having enrolled in law school of her own choice or because she had followed her family's advice.

Results

As preregistered, we tested the proportion of participants exhibiting a preference for agency over the negative outcomes against 50% with chi-square tests. In all three scenarios, the majority of participants preferred being the agent of the negative outcome over having somebody else causing it. Specifically, in the car scenario (N = 149), more participants (75.2%) preferred having gotten into a car accident while they rather than a friend were driving (χ^2 (1) = 37.75, p < .001; see Fig. 1). In the ski accident scenario (N = 150), 71.3% preferred having lost control of their skis and broken their leg because they had accidentally slipped on an ice sheet rather than somebody else had accidentally hit them (χ^2 (1) = 27.31, p < .001). Finally, in the law school scenario (N = 150), 78.7% thought Emily would prefer dropping out of law school after having chosen to study law herself rather than having followed her family's advice (χ^2 (1) = 49.31, p < .001).[17]

Discussion

The results of study 1 provide initial evidence that people can prefer owning a negative outcome rather than holding others responsible for their misfortune. One may argue that the preference for being responsible for one's own misfortune is not driven by a general desire to own negative outcomes but rather by disliking to blame others for it, especially when these

[17] The aggregate analysis of the three scenarios found that overall our participants (N = 449) indicated an unambiguous preference for receiving a negative outcome when they themselves caused it (75.1%) than when someone else did (χ^2 (1) = 112.75, p < .001).

others are close friends as in the car accident (i.e., blaming one's friend) and the law school dropout scenario (Emily blaming her family). However, this explanation could not account for the observed preferences in the ski accident scenario in which participants preferred being responsible for breaking their own leg rather than blaming an unknown stranger for that outcome.

The ski accident outcome—one may further argue—was not so much due to a lack of skill but more a matter of bad luck; even very seasoned skiers accidentally hit ice shields and fall. So maybe it is not so much that people prefer owning negative outcomes but rather that they are reluctant to blame others for their bad luck.

To test this possibility, in study 2 we chose a scenario in which the outcome is unambiguously determined by skill and knowledge—one's own skill or knowledge or that of a stranger. Furthermore, we added a third option in which the negative outcome was caused by pure chance (i.e., bad luck). According to self-determination preferences, participants should prefer agency over the negative outcome. According to the self-serving preferences, people should prefer not being responsible for negative outcomes because they want to maintain a positive self-image. They should hence be happy to attribute the negative outcome to a stranger or to bad luck rather than to themselves.

STUDY 2

Participants in study 2 were asked to answer a difficult probability question by choosing the correct answer out of 10 answer options. They were further told to imagine that, if the selected answer was correct, they would receive a reward of $50. Participants then indicated what they would prefer in case their selected answer was wrong: to have selected the wrong

answer themselves, to have a stranger select the wrong answer on their behalf, or to have a random device select the wrong answer on their behalf (the probability of that random device choosing the wrong answer was hence 90%). Apart from assessing participants' agency preferences, we also asked them who would be more likely to get the probability question right, they themselves or the stranger. This measure allowed us to test whether overplacement may be driving agency preferences.

Method

Participants. Three hundred three Amazon Mechanical Turk workers from the US with a minimum approval rating of 95% completed the experiment in exchange for monetary compensation ($0.20 for 2 minutes expected completion time). As preregistered, we excluded the responses of 71 participants who failed to correctly answer an attention-check question, our final sample hence consisted of 232 responses (50% female, M_{Age} = 36.6 years, SD_{Age} = 10.1 years).

Procedure. Study 2 employed a 3 (agent: self vs. other vs. chance) within-subject design. Participants were asked the following probability question (adapted from Vosgerau, 2010): "When rolling a fair 6-sided die four times, what is the probability of tossing a "3" at least once?" Without knowledge of probability theory, this question is difficult to answer, so we expected most participants would not be able to answer it correctly, or at least would be uncertain about their ability to do so. Participants were told that one of the following 10 answer options was correct:

(1) 4/625

(2) 1/625

(3) 54/346

(4) 4/6

(5) 16/36

(6) 817/3014

(7) 671/1296

(8) 425/863

(9) 3/16

(10) 625/1296

The correct answer is (7) which was chosen by 3.9% of participants. Next—without receiving feedback about the correctness of their chosen answer—participants were asked to imagine facing the same question with one difference: this time, the reward for answering correctly would be $50. Participants indicated which of the following would be their preferred way of not winning the $50: (1) choosing a wrong answer themselves, (2) having someone else choose a wrong answer on their behalf, or (3) having a computer randomly pick a wrong answer on their behalf. Finally, participants indicated whom they thought was more likely to answer the question correctly—they or someone else—using a 9-point unnumbered scale (definitely that other person-definitely myself).

Results

As preregistered, we first tested whether all three preference options were chosen equally often and found they were not (χ^2 (2) = 49.56, $p < .001$). More participants (50.9%) preferred answering the question incorrectly themselves over someone else (13.4%; χ^2 (1) = 50.80, $p < .001$) or a computer (35.8%; χ^2 (1) = 6.09, $p = .014$; see Fig. 2) answering it incorrectly on their behalf.

To test whether overplacement—participants' belief that they are more skilled/knowledgeable than others—is driving these results, we analyzed participants' responses to the question asking who—they or someone else—was more likely to answer the probability question correctly (-4 = *definitely that other person*, 0 = *same probability*, 4 = *definitely myself*). A one-sample t-test showed the mean (M = 0.66, SD = 2.03) to be significantly greater than 0 ($t(231) = 4.96$, $p < .001$; cf., Fig. 3), suggesting that, overall, participants thought they had a slightly better chance of picking the right answer than someone else picking it on their behalf.

Since the subsequent overplacement-related analyses were not preregistered, we analyzed the data in three different ways to probe the robustness of the results. All three analyses show the same result.

First, a logistic regression of participants' counterfactual preferences on their rated perceived likelihood of answering the probability question correctly did not yield a significant effect ($b = 0.02$, $z = 0.32$, $p = .748$), suggesting overplacement is not underlying the observed preference for agency over the prospect of not winning the $50 reward.

Second, we divided the sample (N = 232) into two subsamples representing those who believed they were more likely to solve the question correctly (ratings greater than 0 on the confidence scale; N = 96) and those who did not (ratings of 0 or less on the confidence scale; N = 136), and ran the main analysis on choice (i.e., chi-squared tests) for each group. If overplacement was at play, we should replicate our findings with responses from the overplacement subsample but not with responses from the no-overplacement subsample. As shown in Figure 4, we replicated our findings for both groups. In the overplacement group, the proportion of participants who preferred not winning $50 by answering themselves (50%) was larger than the proportion of those who preferred having someone else reach the same outcome

on their behalf (15.6%; $\chi^2(1) = 17.29$, $p < .001$), and marginally larger than the proportion of those who preferred having a computer do so on their behalf (34.4%; $\chi^2(1) = 2.78$, $p = .096$). Responses from the no-overplacement subsample revealed the same pattern: the proportion of participants who preferred to answer the question incorrectly by themselves (51.5%) was larger than the proportion of those who preferred having someone else reach the same outcome on their behalf (11.8 %; $\chi^2(1) = 33.91$, $p < .001$), and marginally larger than the proportion of those who preferred having a computer do so on their behalf (36.8%; $\chi^2(1) = 3.33$, $p = .068$).

Lastly, we repeated the same analyses dividing our sample into three subsamples: participants who believed they were more likely (ratings > 0; N = 96), as likely (ratings = 0; N = 95), and less likely (ratings < 0; N = 41) than others to answer the question correctly. As shown in Figure 5, this approach generated a similar pattern across groups. In the overplacement group, the proportion of participants who preferred not winning $50 by answering themselves (50%) was larger than the proportion of those who preferred having someone else reaching the same outcome on their behalf (15.6%; $\chi^2(1) = 17.29$, $p < .001$), and marginally larger than the proportion of those who preferred having a computer do so on their behalf (34.4%; $\chi^2(1) = 2.78$, $p = .096$). In the "same probability" group the proportion of participants who preferred not winning $50 by answering themselves (49.5%) was larger than the proportion of those who preferred having someone else reaching the same outcome on their behalf (12.6%; $\chi^2(1) = 21$, $p < .001$), but not significantly larger than the proportion of those who preferred having a computer do so on their behalf (37.9%; $\chi^2(1) = 1.5$, $p = .227$). Finally, in the underplacement group, the proportion of participants who preferred not winning $50 by answering themselves (56.1%) was larger than the proportion of those who preferred having someone else reaching the same outcome on their behalf (9.8%; $\chi^2(1) = 13$, $p < .001$), but not significantly larger than the

proportion of those who preferred having a computer do so on their behalf (34.2%; χ^2 (1) = 2.2, p = .139).

Discussion

The results of study 2 provide further evidence of people's preference for personal agency over negative outcomes. Participants preferred owning a negative outcome not only over somebody else causing it, but also over pure chance (i.e., bad luck) causing it, suggesting that this preference constitutes a general preference for internal over external attribution of negative outcomes. Also, this was true even if the outcome was clearly the result of one's skill/knowledge, ruling out the possibility that people simply dislike blaming others for their misfortune. Finally, this preference cannot be explained by overplacement—peoples' belief that they are more skilled/knowledgeable than others—since it was observed in all groups, those who overplaced themselves, those who did not, and those who underplaced themselves.

STUDY 3

Study 3 was designed to test whether the preference for agency over negative outcomes observed in studies 1 and 2 can be explained by the impact bias. Research on affective forecasting shows people tend to overestimate affective intensity when forecasting the impact of future events and experiences (e.g., Gilbert et al., 1998; Gilbert et al., 2004a; Wilson & Gilbert, 2003), especially when considering atypical events (Gilbert et al., 2004b), arguably because atypical experiences of the past are more memorable than typical ones (Morewedge, Gilbert, & Wilson, 2005). For the impact bias to qualify as driving our effect, instances of negative outcomes caused by others and chance would need to be perceived as more atypical than those

caused by oneself, and therefore should be easier to recall and induce stronger negative affect. In this case, people may be motivated to attribute negative outcomes to themselves to minimize negative affect.

To test this hypothesis, in study 3 we asked participants to recall instances of experiencing a negative outcome that was caused either by themselves, by somebody else, or by chance. They then rated the recalled instances for ease of retrieval and affective intensity, so we could test whether negative outcomes caused by oneself are less affectively intense than negative outcomes caused by others and chance.

Method

Participants. One hundred and one Amazon Mechanical Turk workers from the US with a minimum approval rating of 95% completed the experiment in exchange for monetary compensation ($0.30 + $0.50 bonus for 5 minutes expected completion time). We incentivized participants to provide detailed accounts of past events by offering a $0.50 bonus for thoughtful, detailed responses.[18] We preregistered to exclude data from participants who failed to follow the instructions (i.e., did not recall all three events as instructed). Two research assistants blind to the experimental hypotheses independently coded responses to ensure the events they recalled fit the relevant experimental conditions (i.e., negative outcome caused by oneself, someone else, and chance). The two research assistants agreed that 69 participants (46.4% female, 0% other, M_{Age} = 38.4 years, SD_{Age} = 12.4 years) had provided suitable descriptions for all three recalled events.

Procedure. Participants were told the experimenters were interested in learning about experiences from their past that involved behaviors resulting in negative outcomes. They were

[18] Thirteen of 101 participants did not receive the bonus, because their answers were either meaningless or not at all detailed.

then asked to provide a detailed description of three events: one in which they experienced a negative outcome caused by their own actions, one in which they experienced a negative outcome caused by someone else, and one in which they experienced a negative outcome caused by chance. After recalling each event, participants indicated how long ago the event happened by choosing one of five options ranging from 1 (*in the last month*) to 5 (*more than 5 years ago*), and how difficult retrieving the details of the event was (1 = *not at all difficult*, 10 = *very difficult*). Participants were then asked to come up with a short title for each event. At the end of the study, we showed participants the three titles they had assigned to their events and asked them to indicate how negative each recalled event was (1 = *not at all negative*, 7 = *very negative*). The study concluded with demographic questions.

Results

Recalled negative outcomes caused by oneself occurred marginally more recently (M_{Self} = 3.54, SD_{Self} = 1.36) than those caused by someone else (M_{Other} = 3.87, SD_{Other} = 1.15; $t(68) = -1.98$, $p = .052$), but were experienced at about the same time as negative outcomes caused by chance (M_{Chance} = 3.32, SD_{Chance} = 1.36; $t(68) = 1.18$, $p = .243$).

Pair-wise comparisons show that the recalled outcomes caused by participants themselves were more negative (M_{Self} = 6.09, SD_{Self} = 1.20) than outcomes caused by others (M_{Other} = 5.16, SD_{Other} = 1.43; $t(68) = 5.18$, $p < .001$) and by chance (M_{Chance} = 5.10, SD_{Chance} = 1.59; $t(68) = 4.24$, $p < .001$). Participants further indicated that recalling outcomes they had caused themselves was easier (M_{Self} = 2.93, SD_{Self} = 2.61) than recalling outcomes that others had caused (M_{Other} = 3.72, SD_{Other} = 2.89; $t(68) = -2.80$, $p = .007$), but as difficult to retrieve as outcomes that were caused by chance (M_{Chance} = 3.38, SD_{Chance} = 2.58; $t(68) = -1.57$, $p = .121$; see Fig. 6).

In two separate regressions, we regressed outcome negativity and recall difficulty on outcome cause and the time the outcomes had occurred (we used 4 dummies for the five time categories). Replicating our previous results, outcomes caused by participants were rated more negatively than those caused by others (b = 1.05, t(200) = 4.60, p < .001) and by chance (b = 0.91, t(200) = 3.92, p < .001). However, recall difficulty of outcomes caused by oneself was not different from recall difficulty of outcomes caused by others (b = -0.79, t(200) = -1.68, p = .094) and by chance (b = -0.39, t(200) = -0.82, p = .414; see Table 2.1).

Discussion

The results of study 3 are inconsistent with the impact bias driving a preference for agency over negative outcomes. If the impact bias were underlying this preference, negative outcomes caused by others or chance should be less common than those caused by participants and should be easier to retrieve and induce greater negative affect. The results of study 3 show the opposite—negative outcomes caused by others were generally more difficult to retrieve and induced less negative affect than negative outcomes caused by oneself.

Having found no evidence for overplacement and the impact bias driving the preference for agency over negative outcomes, in the remaining studies we focus on testing our own account for why and when such preference occurs. We hypothesize that, motivated by a desire for self-determination, people exhibit a general preference for agency when outcomes are exclusively attributable to a sole agent. In other words, if people are asked to choose between attributing a negative outcome either to themselves or to somebody else, they prefer having agency and assuming responsibility for the outcome. However, when outcomes are jointly caused by

multiple agents, people's desire for agency is already satisfied, so they can attend to their second motive to protect their self-esteem by assuming some agency over the outcome but attributing a larger share of it to others. In studies 4A, 4B, and 5 we test our hypothesis by manipulating the relative saliency of agents and outcomes. Specifically, we test whether the preference for agency over negative outcomes becomes weaker when participants' attention is drawn away from agents towards outcomes.

Because we did not observe differences between the self–other and self–chance comparisons in studies 2 and 3, the following studies include only negative outcomes caused by participants and by others.

STUDY 4A

According to the isolation effect in the elimination by aspects theory (Tversky, 1972), individuals facing a choice task tend to isolate (i.e., focus on) the choice aspect/characteristic that is most important to them, according to a hierarchy of motives. In the context of our investigation, in which participants compared agent scenarios caused exclusively by a sole agent (either themselves or somebody else), we predict that the most important aspect participants isolate is self-determination. Since the two choice options (i.e., receiving a negative outcome because of oneself or because of someone else) vary by self-determination, people are able to choose the option that grants them agency over their outcome. For the isolation effect to occur, individuals need to directly compare choice options—in other words, they need to evaluate choice options jointly (e.g., Hsee, 1996). When agent scenarios are evaluated separately, in fact, all aspects of the scenarios are distinct, and they are harder—or impossible—to isolate. Hence, the preference for agency over negative outcomes should be attenuated. Studies 4A and 4B test

these predictions using different negative outcomes.

In study 4A, we tested preferences for agency in separate and joint evaluations with the same outcome used in study 2.

Method

Participants. Four hundred fifty-one Amazon Mechanical Turk workers from the US with a minimum approval rating of 95% completed the experiment in exchange for monetary compensation ($0.30 for 2.5 minutes expected completion time). We preregistered to exclude responses from participants who failed to correctly answer an attention-check question, resulting in seven exclusions. Our final sample consisted of 444 participants (44.4% female, 1.6% other, M_{Age} = 38.9 years, SD_{Age} = 11.8 years).

Procedure. Participants were randomly assigned to one of three experimental conditions (evaluation mode: joint vs. self vs. other). After answering the probability question, with no feedback given about the correctness of their answers, participants imagined facing the same question but with a $5 bonus for answering correctly. Participants in the "self" condition predicted how dissatisfied[19] they would be if they had answered the question incorrectly; participants in the "other" condition predicted how dissatisfied they would be if someone else answered incorrectly on their behalf. Participants assigned to the joint evaluation condition predicted their dissatisfaction for both cases, having incorrectly answered the question themselves and somebody else having incorrectly answered on their behalf. Dissatisfaction in all conditions was expressed on a 7-points scale (1 = *not at all dissatisfied*, 7 = *very dissatisfied*).

[19] We used predicted dissatisfaction (instead of preference) as a dependent variable because in separate evaluation participants can not express a preference.

Finally, participants answered an attention-check question and provided demographic information.

Results

First, we tested whether the exclusions caused differences in age and gender across the two experimental conditions. Before exclusions, participants did not differ in age ($F(2, 448) = 0.08$, $p = .927$) or gender ($\chi^2(4) = 4.33$, $p = .363$) across conditions. This finding held after exclusions: participants did not differ in age ($F(2, 441) = 0.10$, $p = .909$) or gender ($\chi^2(4) = 4.13$, $p = .389$) across conditions.

Second, we ran the preregistered analyses, comparing dissatisfaction ratings in the joint evaluation condition with a paired t-test and dissatisfaction ratings in the separate evaluation conditions with an independent samples t-test. As hypothesized, in joint evaluation we replicated the preference for agency over negative outcomes, as participants predicted being less dissatisfied when they answered incorrectly on their own ($M_{Joint\text{-}Self} = 4.30$, $SD_{Joint\text{-}Self} = 2.14$) than when someone else answered incorrectly on their behalf ($M_{Joint\text{-}Other} = 4.82$, $SD_{Joint\text{-}Other} = 2.21$; $t(146) = -2.91$, $p = .004$). In separate evaluations, in contrast, dissatisfaction ratings were not statistically different from one another ($M_{Separate\text{-}Self} = 4.57$, $SD_{Separate\text{-}Self} = 1.76$, $M_{Separate\text{-}Other} = 4.68$, $SD_{Separate\text{-}Other} = 1.95$; $t(295) = -0.52$, $p = .604$; see Fig. 7).

Finally—even though not preregistered[20]—we ran a linear mixed-effect model regressing dissatisfaction level on agent, evaluation mode, and their interaction, with participants as random effect. Specifying participants as random effect took care of the fact that ratings in joint evaluations were provided by the same participant whereas ratings in separate evaluations were

[20] This analysis was not preregistered because we realized only after collecting the data how to estimate the interaction effect in this experimental design.

provided by different participants. Even though in the predicted direction, the interaction did not reach significance ($b = 0.41$, $t = 1.42$, $p = .157$; see Table 2.2).

Discussion

The results of study 4A suggest that the preference for agency over negative outcomes is muted when saliency is drawn away from the agents. When participants could directly compare negative outcomes caused by themselves and by somebody else, the agent (isolated, distinctive aspect) was the most salient choice aspect, elimination by agent was possible, and participants preferred agency. In contrast, when participants evaluated the two agency-scenarios separately, no comparison based on agents was possible, and participants no longer showed a preference for one over the other.

While we found empirical support for these predictions with the preregistered analysis, the interaction was not significant. We suspected that our study was underpowered to detect said interaction. To address this point, we ran study 4B with a larger sample size. As negative outcome, we used the ski-accident scenario from study 1 that resembles an affective-rich situation that participants may encounter in their everyday lives.

STUDY 4B

Method

Participants. One thousand forty-nine CloudResearch approved Amazon Mechanical Turk workers from the US completed the experiment in exchange for monetary compensation ($0.25 for 2 minutes expected completion time). We preregistered that participants who failed to correctly answer an attention-check question would not be allowed to start the study. Thus, we

had no exclusions and our final sample consisted of 1,049 participants (56.3% female, 1.1% other, M_{Age} = 39.8 years, SD_{Age} = 12.5 years).

Procedure. As study 4A, study 4B employed a 3 (evaluation mode: joint vs. self vs. other) between-participants design, using the ski accident scenario from study 1. Participants imagined having their right leg badly broken after having lost control of their skis. Participants in the "self" condition predicted how dissatisfied they would be if they had lost control of their skis and broken their leg because they had accidentally slipped on an ice sheet; participants in the "other" condition predicted how dissatisfied they would be if they had lost control of their skis and broken their leg because another skier had accidentally hit them. Participants assigned to the joint evaluation condition predicted their dissatisfaction for both scenarios. Finally, participants provided demographic information.

Results

As preregistered, we ran a linear mixed-effect model regressing dissatisfaction level on agent, evaluation mode, and their interaction, with participants as random effect. That analysis revealed no main effect of evaluation mode (b = 0.04, t = 0.38, p = .700), and a significant main effect of agent (b = 0.73, t = 8.13, p < .001) such that participants reported higher dissatisfaction when someone else caused the negative outcome. Importantly, it yielded a significant interaction showing that—as predicted—the difference in dissatisfaction ratings was larger in joint than in separate evaluations (b = 0.51, t = 3.68, p < .001; see Fig. 8 and Table 2.3). When participants evaluated the two scenarios jointly, they predicted being less dissatisfied breaking their leg when they had accidentally slipped on an ice sheet ($M_{Joint\text{-}Self}$ =5.50, $SD_{Joint\text{-}Self}$ = 1.58) than when another skier had accidentally hit them ($M_{Joint\text{-}Other}$ = 6.23, $SD_{Joint\text{-}Other}$ = 1.26; t(349) = -7.90, p <

.001). When participants evaluated the two scenarios separately, the same pattern was observed but differences in dissatisfaction were attenuated ($M_{Separate\text{-}Self}$ = 6.05, $SD_{Separate\text{-}Self}$ = 1.43, $M_{Separate\text{-}Other}$ = 6.27, $SD_{Separate\text{-}Other}$ = 1.20; $t(697) = -2.26$, $p = .024$).

Discussion

Replicating the pattern observed in study 4A, the findings of study 4B show that people prefer agency over negative outcomes when scenarios are directly compared to each other in joint evaluations, but that preference is attenuated when scenarios are evaluated separately.

STUDY 5

In study 5, we used a different manipulation than joint versus separate evaluations to make the agency aspect more or less salient. We manipulated the salience of the agent by varying the salience of the negative outcome. All participants evaluated two outcomes, a positive and a negative one. We varied the order in which the two outcomes were presented. Our hypothesis was that valuating the negative outcome after the positive one should evoke contrast and increase the salience of the negative outcome at the expense of salience of agency (de Bruin & Keren, 2003). We hence expected the preference for agency over the negative outcome to be attenuated compared to when participants valuated the negative outcome first.

Method

Participants. One thousand seven CloudResearch approved Amazon Mechanical Turk workers from the US and Canada with a minimum approval rating of 95% completed the experiment in exchange for monetary compensation ($0.35 for 3 minutes expected completion

time). We preregistered to exclude responses from participants who failed to correctly answer an attention-check question. After 102 exclusions, our final sample consisted of 905 participants (46.6% female, 0.7% other, M_{Age} = 38.3 years, SD_{Age} = 11.8 years).

Procedure. The study used a 2 (outcome salience: low vs. high; between-participants) x 2 (agent: self vs. other; within-participants) design. As in studies 2 and 4A, participants first answered the probability question "When rolling a fair 6-sided die four times, what is the probability of tossing a "3" at least once?" without being given feedback whether their answer was correct or not. Next, participants imagined facing the same question but with a $5 bonus for answering it correctly. Using the same dependent variable as studies 4A and 4B, we asked half of the participants how satisfied they would be if they answered the question *correctly* and how satisfied they would be if someone else answered it *correctly* on their behalf (1 = *not at all satisfied*, 7 = *very satisfied*). Next, they considered a negative outcome and rated how dissatisfied they would be if they answered the question *incorrectly* and how dissatisfied they would be if someone else answered it *incorrectly* on their behalf (1 = *not at all dissatisfied*, 7 = *very dissatisfied*). The other half of our sample was asked the same questions in reverse order (i.e., starting with the negative outcome). Finally, participants answered an attention-check question and provided some demographic information.

Results

Before proceeding with the main analysis, we tested whether the exclusions had caused differences in age or gender across the two experimental conditions. Before exclusions, participants did not differ in age ($t(1,004.8) = 0.10$, $p = .918$) or gender ($\chi^2(2) = 1.46$, $p = .482$)

across conditions. This finding held after exclusions: participants did not differ in age ($t(902.94) = 0.02$, $p = .987$) or gender ($\chi^2(2) = 2.93$, $p = .231$) across conditions.

Replicating the results of studies 1, 2, 4A and 4B, participants in the "low outcome salience" condition (negative outcome evaluated first) indicated they would be less dissatisfied if they themselves were responsible for the incorrect answer ($M_{Self} = 4.69$, $SD_{Self} = 1.89$) than if someone else had answered incorrectly on their behalf ($M_{Other} = 4.97$, $SD_{Other} = 1.87$; $t(457) = -3.06$, $p = .002$). By contrast, participants in the "high outcome salience" condition (negative outcome evaluated after the positive outcome) expected to be equally dissatisfied with the incorrect answer in both scenarios ($M_{Self} = 4.72$, $SD_{Self} = 2.00$, $M_{Other} = 4.71$, $SD_{Other} = 2.12$; $t(446) = 0.12$, $p = .907$; see Fig. 9). A 2x2 mixed ANOVA[21] on participants' responses to the negative-outcome-scenario ratings revealed no main effect of outcome order ($F(1, 903) = 1.15$, $p = .285$), a significant main effect of agent ($F(1, 903) = 4.22$, $p = .040$) such that participants reported being more dissatisfied when someone else caused the negative outcome, and a significant interaction ($F(1, 903) = 4.93$, $p = .027$) showing that this effect was attenuated in the "high outcome salience" condition.

Discussion

Collectively, the results of studies 4A, 4B and 5 support the proposition that focusing participants' attention away from the agent attenuates the preference for agency over negative outcomes. Studies 4A and 4B showed this by manipulating whether agency-outcome scenarios were evaluated jointly or separately, while study 5 demonstrated this by varying the order in which participants rated agency-outcome scenarios.

[21] To test the interaction, we had preregistered a t-test on the difference in dissatisfaction ratings between the self and the other conditions: $t(1,805.6) = -3.14$, $p = .002$.

In the last two studies, we test our hypothesis that a general preference for agency and self-determination occurs only when outcomes are caused exclusively by a sole agent (oneself *or* someone else). When outcomes are caused jointly by multiple agents (oneself *and* someone else), we predict that people's desire for autonomy is already satisfied and they can protect their self-esteem by assuming some agency over the outcome but attributing a larger share of it to others. This would be the case because when multiple agents jointly cause a negative outcome, the decision-maker has agency anyways. Isolation effect in elimination by aspect theory states that, when the choice options do not vary by the isolated aspect, option elimination is not possible, and the second most important aspect (according to the individual's hierarchy of motives—in our case self-enhancement) is isolated. When asked to evaluate one's own portion of responsibility versus someone else's portion of responsibility for a negative outcome, these two options do vary by self-enhancement. Therefore, elimination by aspect is possible, and the most flattering option (someone else's portion of responsibility) is chosen.

STUDY 6A

In study 6A, we test whether the preference for agency over negative outcomes occurs when outcomes are construed as being exclusively caused by a sole agent but reverses to a preference for self-serving attributions when outcomes are construed as being caused jointly by multiple agents. As a negative outcome, we used again the probability scenario from studies 2, 4A, and 5.

Method

Participants. Four hundred CloudResearch approved Amazon Mechanical Turk workers

from the US completed the experiment in exchange for monetary compensation ($0.25 for 2 minutes expected completion time). We preregistered that participants who failed to correctly answer an attention-check question would not be allowed to start the study. Thus, we had no exclusions and our final sample consisted of 400 participants (54.3% female, 0.3% other, M_{Age} = 40.3 years, SD_{Age} = 12.3 years).

Procedure. Study 6A employed a 2 (negative outcome cause: single agents vs. multiple agents; between-participants) x 2 (agent: self vs other; within-participants) design. Participants were randomly assigned to one of the two between-participants experimental conditions. After answering the probability question, with no feedback, participants imagined facing the same question, but with a $5 bonus for answering correctly. Participants in the "single agent" condition were told that they would have one of two alternative ways to win the bonus: either (1) they would answer the question themselves, and—if their answer was correct—win the $5 bonus, or (2) they would be paired with another participant and—if this participant's answer was correct—win the $5 bonus. Participants then predicted how dissatisfied they would be if they had chosen the wrong answer themselves, and how dissatisfied they would be if someone else had chosen the wrong answer on their behalf. Participants in the "multiple agents" condition were told they would have two attempts to win the bonus: (1) first, they would answer the question themselves, and—if their answer was correct—win the $5 bonus, and (2) then they would be paired with another participant and—if that participant's answer was also correct—win an additional $5 bonus. Participants predicted how dissatisfied they would be if they had chosen the wrong answer themselves, and how dissatisfied they would be if someone else had chosen the wrong answer on their behalf. Finally, participants provided demographic information.

Results

As preregistered, we ran a mixed ANOVA on dissatisfaction ratings that yielded no main effect of either negative-outcome-cause condition ($F(1, 398) = 0.03$, $p = .857$) nor agent ($F(1, 398) = 0.01$, $p = .930$). However, it revealed a significant interaction between negative-outcome-cause condition and agent ($F(1, 398) = 11.84$, $p < .001$; see Fig. 10). As predicted—and replicating our previous results—in the "single agent" condition participants predicted being less dissatisfied for not winning the $5 bonus if they had chosen the wrong answer themselves ($M_{Single\text{-}Self} = 4.41$, $SD_{Single\text{-}Self} = 1.86$) than when someone else had chosen the wrong answer on their behalf ($M_{Single\text{-}Other} = 4.78$, $SD_{Single\text{-}Other} = 1.86$; $t(198) = -2.40$, $p = .017$). In contrast, in the "multiple agents" condition the opposite pattern was observed: participants predicted being more dissatisfied for not winning the $5 bonus if they had chosen the wrong answer themselves ($M_{Multiple\text{-}Self} = 4.75$, $SD_{Multiple\text{-}Self} = 1.93$) than when someone else had chosen the wrong answer on their behalf ($M_{Multiple\text{-}Other} = 4.39$, $SD_{Multiple\text{-}Other} = 1.93$; $t(200) = 2.50$, $p = .015$).

Discussion

Attributional preferences were reversed in study 6A, demonstrating a preference for both internal and external attribution for the same negative outcome depending on whether the outcome was construed as being caused by a sole agent exclusively or by multiple agents jointly. When a negative outcome was caused by a sole agent, participants preferred assuming agency as a means to satisfy their desire for self-determination. However, when a negative outcome was caused jointly by multiple agents, participants preferred self-serving attributions.

While study 6A provides empirical support for a reversal of attributional preferences

using the probability question paradigm, in study 6B we chose a classical paradigm from the self-serving attribution literature, Larson's (1977) two students working together on a 30-minutes problem-solving task.

STUDY 6B

Larson (1977) asked 112 male students to work together in pairs on a 30-minutes task consisting of solving cryptograms, word puzzles featuring encrypted text that participants are asked to decrypt to reveal a message of some sort. One third of the student pairs was assigned to the success condition in which they had to solve 6 cryptograms and were given feedback that their performance was above average. Another third of student pairs was assigned to the failure condition, these students had to solve 6 cryptograms and were told that they performance was below average. The remaining third of student pairs was assigned to the control condition and could solve cryptograms ad libitum without feedback. The results showed self-serving attributions in the negative domain. Specifically, students in the failure condition attributed more responsibility to their partners than to themselves. They also attributed less responsibility to themselves in the failure condition than in the control and success conditions.

In study 6B, we created a scenario based on this paradigm in which two students, Jamie and Logan, worked on a problem-solving task. We asked participants to predict how dissatisfied Jamie would be if the group project failed, either because Jamie, Logan, or both jointly had made a fatal mistake.

Method

Participants. Three hundred ninety-eight CloudResearch approved Amazon Mechanical

Turk workers from the US completed the experiment in exchange for monetary compensation ($0.25 for 2 minutes expected completion time). We preregistered that participants who failed to correctly answer an attention-check question would not be allowed to start the study. Thus, we had no exclusions and our final sample consisted of 398 participants (58.0% female, 1.0% other, M_{Age} = 39.2 years, SD_{Age} = 12.0 years).

Procedure. As study 6A, study 6B employed a 2 (negative outcome cause: single agent vs. multiple agents; between-participants) x 2 (agent: self vs other; within-participants) design. Participants imagined two college students, Jamie and Logan, being assigned to solve a 30-minutes problem-solving task and were told that this task meant a lot to Jamie as their scholarship in the next year depended on it. Participants were further informed that Jamie and Logan's performance on the task turned out to be below average because they had committed a fatal mistake. As a consequence, Jamie did not qualify for the scholarship. At this point, participants were randomly assigned to one of the two between-participants conditions. Those in the "single agent" condition were told that Jamie and Logan had been required to split up the work (i.e., each had been working individually on half of the task). Hence, **only one of them had caused the fatal mistake**. Participants were asked to put themselves in the shoes of Jamie and predict how dissatisfied they would be with: (1) having lost the scholarship because they had caused the fatal mistake themselves, and (2) having lost the scholarship because Logan had caused the fatal mistake. Participants in the "multiple agents" condition were told that Jamie and Logan had been required to work together on the task, so **both had contributed to the fatal mistake**. Participants then predicted how dissatisfied they would be if they were Jamie with: (1) their own contribution to the fatal mistake that led to the scholarship loss, and (2) Logan's

contribution to the fatal mistake that led to the scholarship loss. Finally, participants provided demographic information.

Results

As preregistered and as in study 6A, we ran a mixed ANOVA on dissatisfaction ratings which revealed a main effect of outcome cause condition ($F(1, 396) = 5.42$, $p = .020$) indicating that participants predicted higher dissatisfaction levels in the "single agent" than "multiple agents" condition, and a main effect of agent ($F(1, 396) = 6.81$, $p = .009$) showing that participants reported more dissatisfaction when Jamie (rather than Logan) was the agent. More importantly, the predicted interaction of negative-outcome-cause condition and agent was observed ($F(1, 396) = 35.10$, $p < .001$; see Fig. 11). In the "multiple agents" condition, we replicated the results of study 6A and Larson's (1977) self-serving attribution result, as participants predicted being more dissatisfied with their own (Jamie's) contribution to the fatal mistake ($M_{Multiple\text{-}Self} = 6.04$, $SD_{Multiple\text{-}Self} = 1.36$) than with Logan's contribution to the fatal mistake ($M_{Multiple\text{-}Other} = 5.32$, $SD_{Multiple\text{-}Other} = 1.55$; $t(199) = 6.1$, $p < .001$). In the "single agent" condition, attributional preferences reversed and we observed again a preference for agency over negative outcomes as participants predicted to be more dissatisfied with Logan exclusively causing the fatal mistake ($M_{Single\text{-}Other} = 6.10$, $SD_{Single\text{-}Other} = 1.41$) than with themselves (Jamie) exclusively causing the mistake ($M_{Single\text{-}Self} = 5.82$, $SD_{Single\text{-}Self} = 1.58$; $t(197) = -2.30$, $p = .021$).

Discussion

Using a similar paradigm—albeit hypothetical—as Larson (1977) did for demonstrating self-serving attributions, study 6B shows that attributional preferences reverse when negative

outcomes are exclusively attributable to a single agent. When a negative outcome is attributable only to a single agent, people prefer assuming agency due to their desire for self-determination and autonomy. However, when negative outcomes are caused jointly by multiple agents, people can satisfy their desire for self-determination and at the same time protect their self-esteem and positive self-image by attributing a larger share of responsibility to others.

GENERAL DISCUSSION

Judge Schroeder was right. Had Rittenhouse been convicted, he would have felt better having picked the lottery numbers for alternate jurors himself. This would be the case, because the alternate jurors could have been picked *either* by the defendant *or* by someone else exclusively.

Seemingly contrary to the prediction of a self-serving bias in attribution (e.g., Heider, 1958; Larson, 1977), we show people prefer agency over negative outcomes when the outcome can be attributed exclusively to a single agent. Such agency preferences over negative outcomes cannot be explained by overplacement (e.g., Moore & Healy, 2008), because most participants exhibit them regardless of whether they overplace themselves. Furthermore, the impact bias (Morewedge, Gilbert & Wilson, 2005) cannot account for these preferences, because negative outcomes that one causes are more memorable and more emotionally aversive than negative outcomes that others cause. Instead, studies 4A, 4B and 5 suggest the agency preference is attenuated when people's focus is drawn away from the agents, as isolation and elimination by agency is harder to achieve. Finally, studies 6A and 6B reconcile these findings with the literature on self-serving attributions, showing that self-determination versus self-serving preferences are driven by whether the outcome is construed as caused exclusively by a single

agent or jointly by multiple agents.

The fact that agency preference over negative outcomes mostly occurs in joint evaluations suggests the preference may be prediction error. Real-world outcomes are usually experienced in separate evaluation mode—facing one outcome, typically caused by one individual, at a time. Moreover, *experiencing* an outcome is an affect-rich experience, causing individuals to allocate a larger share of their attention toward the outcome (Buechel, Zhang, Morewedge & Vosgerau, 2014), and we find the agency preference over negative outcomes is attenuated when saliency is drawn away from the agency aspect. This further suggests such preference might constitute a prediction error. Whether the preference is based on a prediction error seems to be an interesting avenue for future research.

Note that all of our studies are hypothetical. This is because, as mentioned above, outcomes exclusively caused by sole agents only happen in one way in real life. For example, people have a car accident either because they were driving *or* because someone else was driving. They can never have the same accident in two ways. Our results are still significant for real-life decision making because preferences are typically formed hypothetically in joint evaluation (i.e., through the generation of counterfactuals). These preferences subsequently influence the choices individuals make. For example, Judge Schroeder made a real, very consequential choice—as it could have served as a precedent for future trials—based on this type of counterfactual thinking.

In conclusion, this paper provides three main contributions. First, it shows that—seemingly contradicting self-serving preferences—a preference for personal agency over negative outcomes exists. Second, it shows it is due to people's desire for self-determination and

autonomy. Third, it reconciles self-determination and self-serving preferences. Specifically, a preference for personal agency over negative outcomes occurs when negative outcomes are attributable exclusively to a sole agent (i.e., oneself *or* someone else), and a preference for self-serving attributions occurs when negative outcomes are caused jointly by multiple agents (i.e., oneself *and* someone else). Also, our results are relevant for real-life decision-making because preferences are typically construed in joint evaluation through hypothetical counterfactual thinking. These preferences then influence people's choices, like Judge Schroeder's choice to let Rittenhouse randomly pick the alternate jurors himself.

ACKNOWLEDGEMENTS

The authors thank Gil Appel, Uma R. Karmarkar, M. Patrick Hulme, Mohamed Hussein, and Elisa Solinas for helpful comments. We also thank Harrison Oliphant and Carolina Raffaelli for excellent research assistance.

Chapter 2, in full, has been submitted for publication of the material. Maimone, Giulia, Joachim Vosgerau and Ayelet Gneezy. The dissertation author is the primary investigator and author of this paper.

DATA AND MATERIALS AVAILABILITY

All preregistrations, data, analyses code, and experimental materials are available on ResearchBox (https://researchbox.org/532&PEER_REVIEW_passcode=EZYJTY).

FIGURES

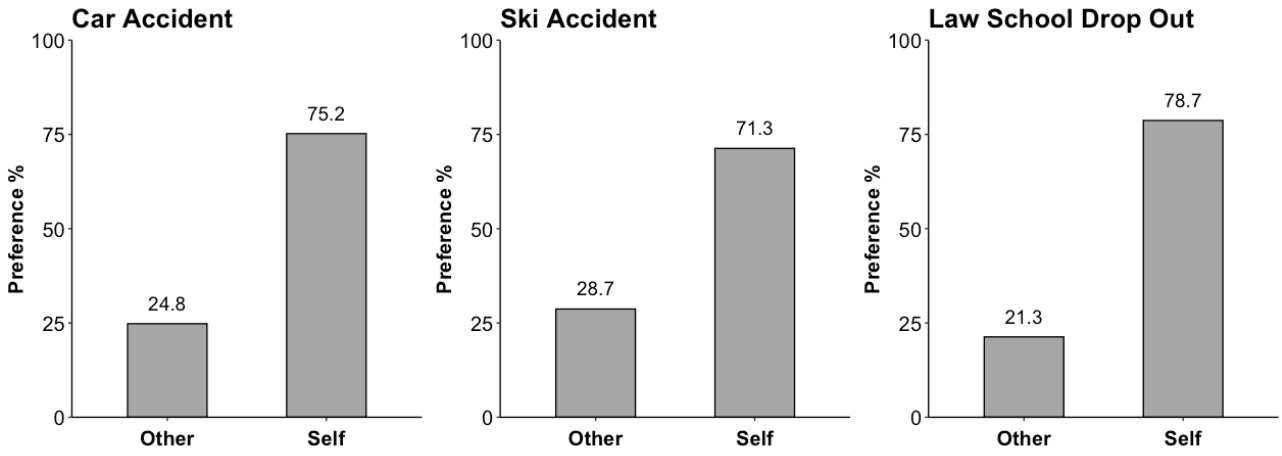


Figure 2.1. Preference for agency (self vs. other) over negative outcomes in the three scenarios in study 1.

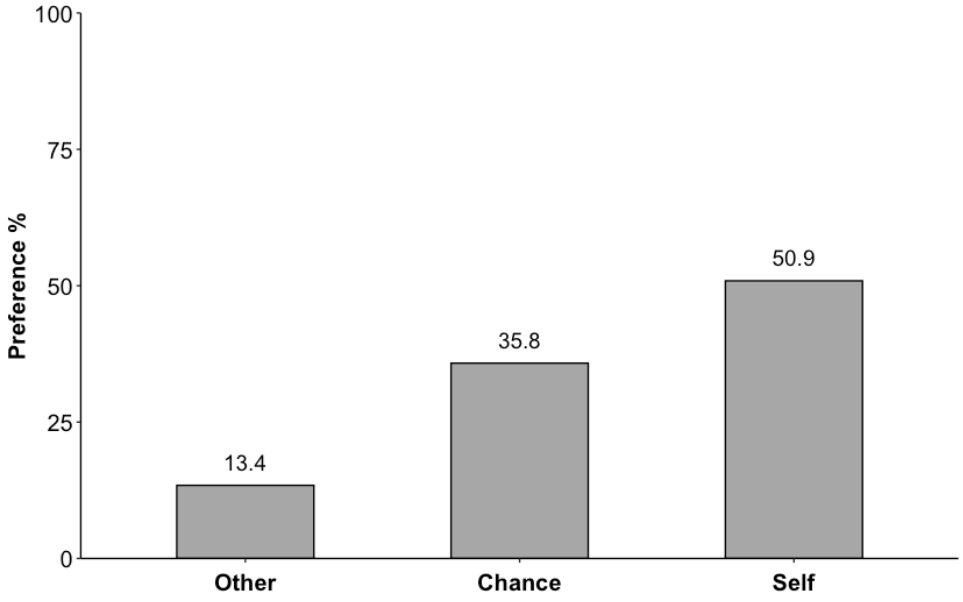


Figure 2.2. Preference for causes (self vs. other vs. chance) of a negative outcome in study 2.

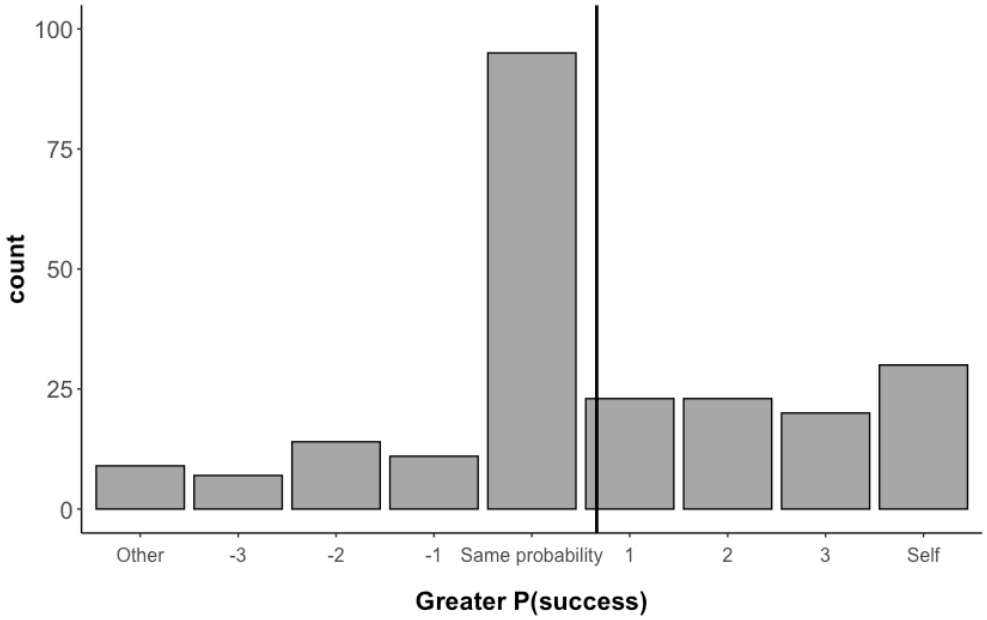


Figure 2.3. Histogram of perceived likelihood of answering the probability question correctly (black line indicates the mean) in study 2.

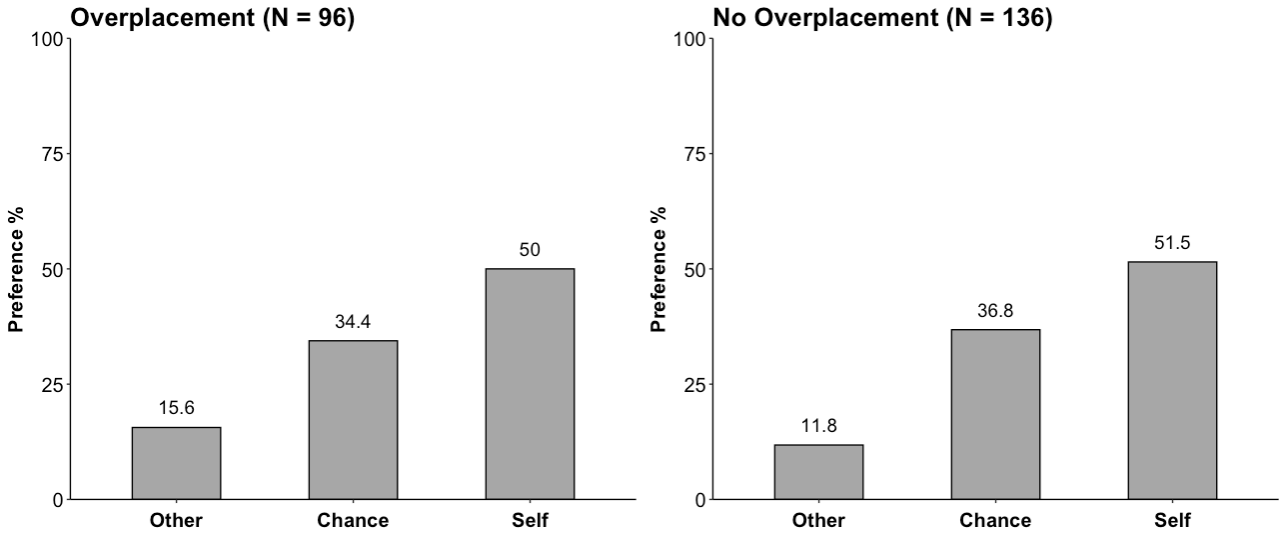


Figure 2.4. Preference for causes (self vs. other vs. chance) of a negative outcome by overplacement (yes vs. not) in study 2.

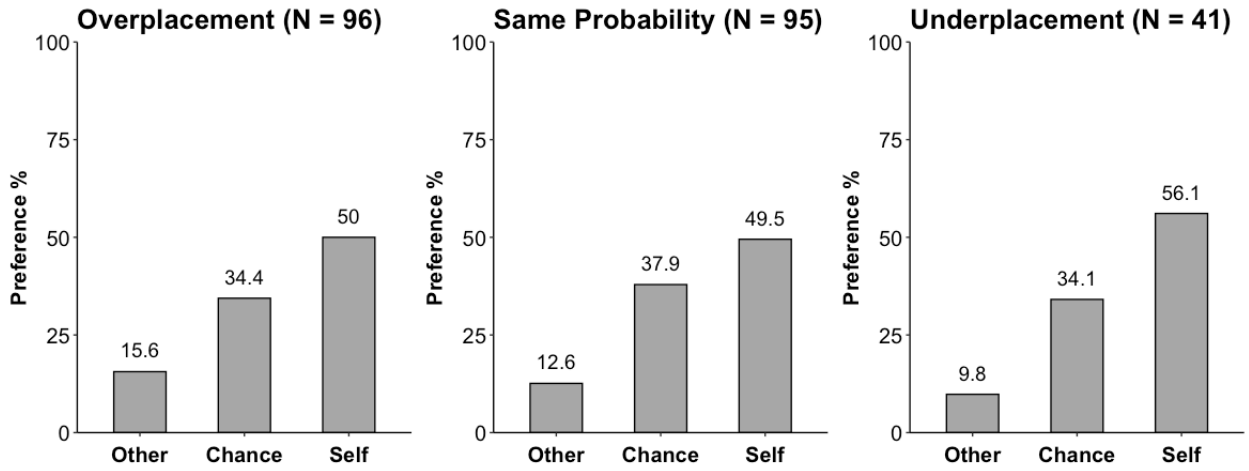


Figure 2.5. Preference for causes (self vs. other vs. chance) of negative outcome by overplacement (overplacement vs. neither vs. underplacement) in study 2.

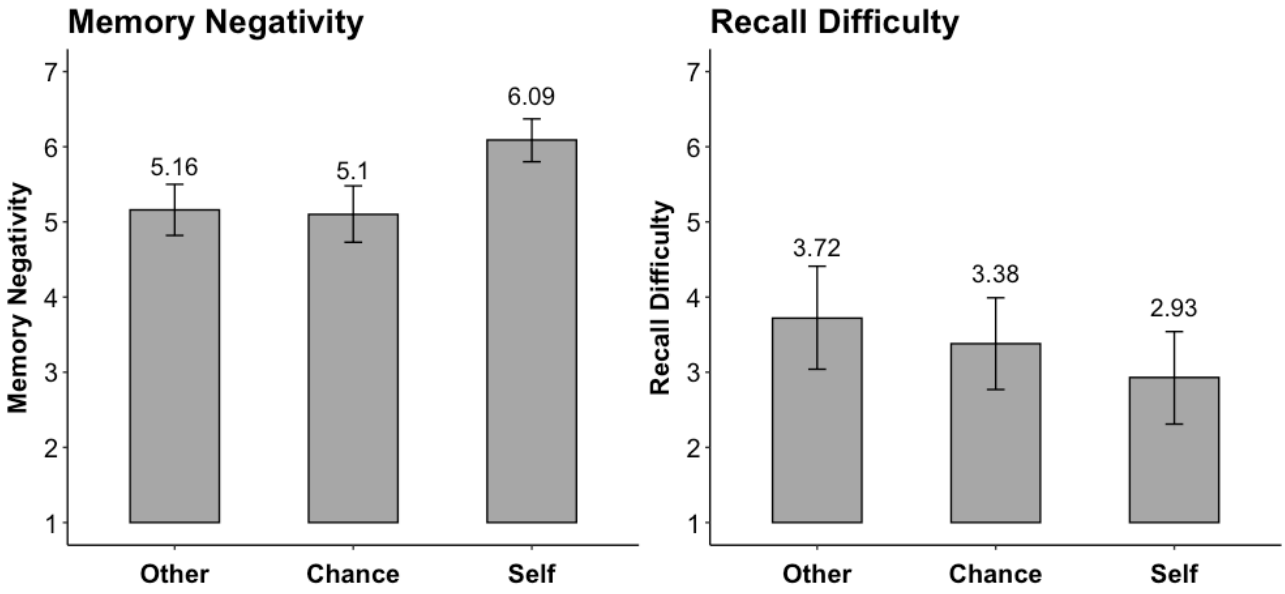


Figure 2.6. Negativity scores (left panel) and recall difficulty (right panel) of negative outcomes by cause (oneself, by someone else, or chance) in study 3. Error bars reflect 95% CIs.

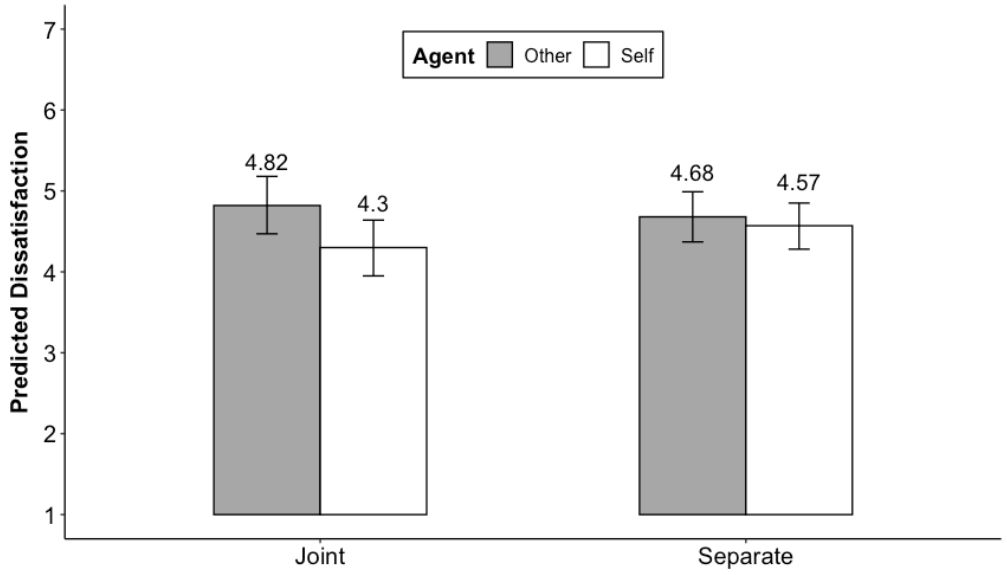


Figure 2.7. Predicted dissatisfaction by evaluation mode and cause of outcome in study 4A. Error bars reflect 95% CIs.

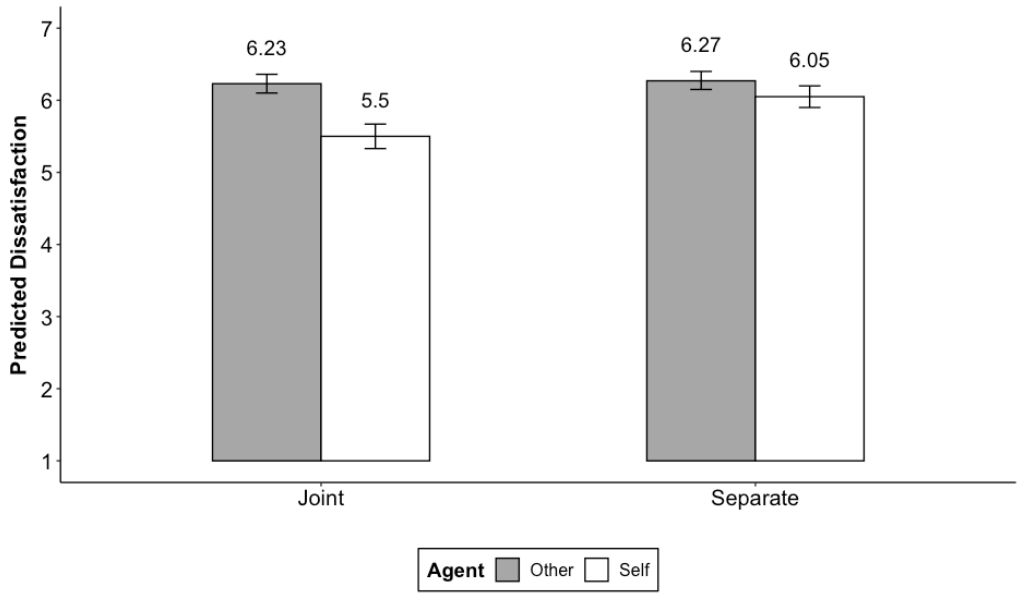


Figure 2.8. Predicted dissatisfaction by evaluation mode and cause of outcome in study 4B. Error bars reflect 95% CIs.

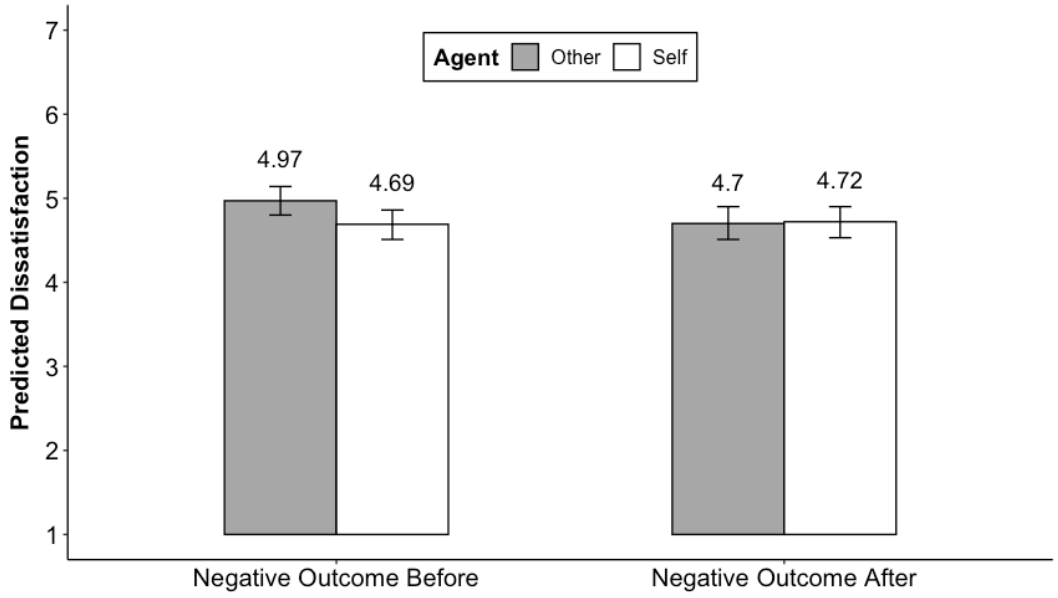


Figure 2.9. Predicted dissatisfaction by outcome salience (low vs. high) and cause (self vs. other) of the negative outcome in study 4. Error bars reflect 95% CIs.

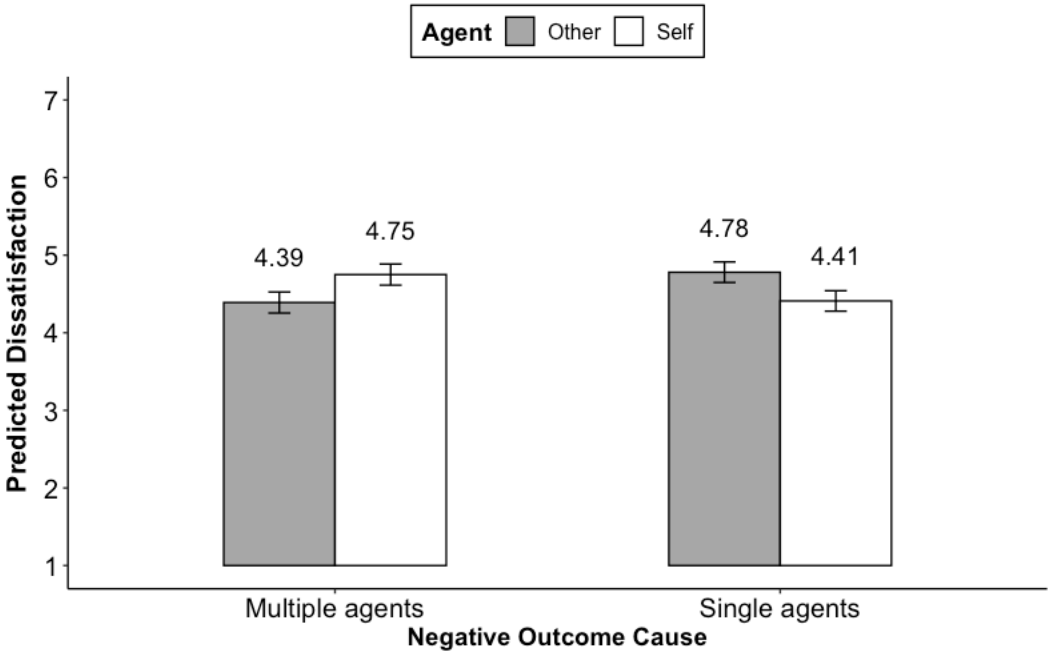


Figure 2.10. Predicted dissatisfaction as a function of agency and number of agents in study 6A. Error bars reflect 95% CIs.

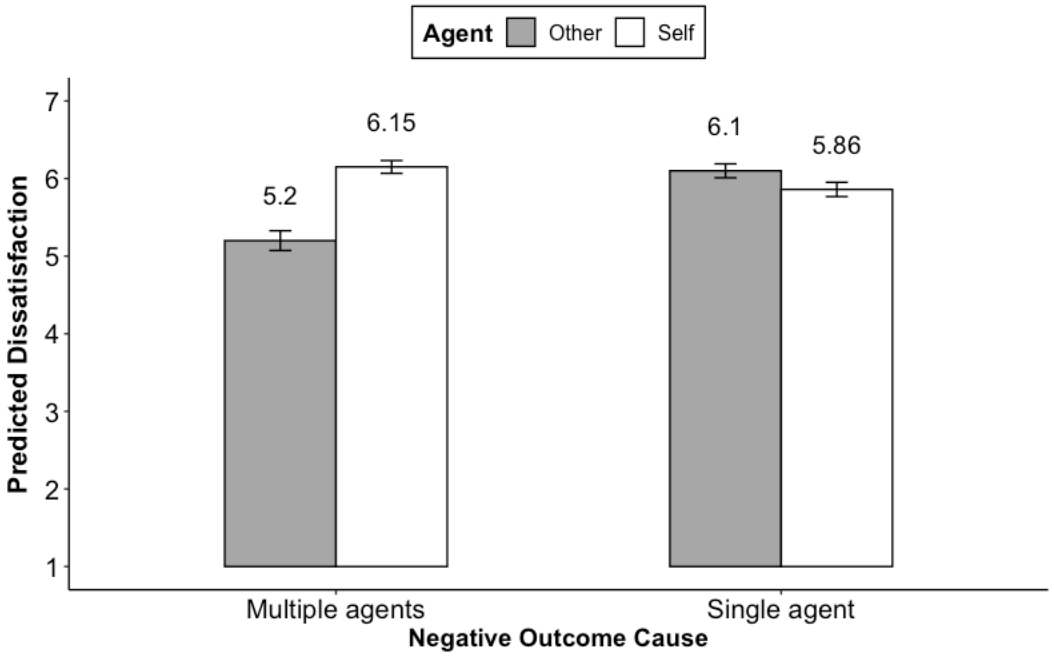


Figure 2.11. Predicted dissatisfaction as a function of agency and number of agents in study 6B. Error bars reflect 95% CIs.

TABLES

Table 2.1. Models regressing negativity scores and recall difficulty on outcome cause (self vs. other vs. chance) and time of the outcome's occurrence (in the last month, in the last 6 months, in the last year, in the last 5 years, more than 5 years ago) in study 3.

	Negativity			**Difficulty**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	5.13	4.52 – 5.74	**<0.001**	3.04	1.79 – 4.29	**<0.001**
Scenario [Chance]	-0.91	-1.36 – -0.45	**<0.001**	0.39	-0.55 – 1.32	0.414
Scenario [Other]	-1.05	-1.51 – -0.60	**<0.001**	0.79	-0.14 – 1.72	0.094
Time [2]	0.22	-0.60 – 1.03	0.601	-0.31	-1.97 – 1.36	0.716
Time [3]	0.69	0.00 – 1.39	**0.049**	0.11	-1.31 – 1.52	0.882
Time [4]	1.22	0.58 – 1.86	**<0.001**	-0.04	-1.36 – 1.29	0.958
Time [5]	1.51	0.85 – 2.16	**<0.001**	-0.24	-1.59 – 1.10	0.719
Observations	207			207		
R^2 / R^2 adjusted	0.218 / 0.194			0.017 / -0.012		

Table 2.2. Model regressing predicted dissatisfaction scores on evaluation mode (joint vs. separate), outcome cause (self vs. other), their interaction, and subjects as random effects in study 4A.

	Dissat		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.82	4.50 – 5.15	**<0.001**
Agent [Self]	-0.52	-0.86 – -0.18	**0.003**
EvaluationMode [S]	-0.15	-0.60 – 0.31	0.531
Agent [Self] * EvaluationMode [S]	0.41	-0.16 – 0.98	0.157
Random Effects			
σ^2	2.21		
$\tau_{00\ ID}$	1.81		
ICC	0.45		
$N_{\ ID}$	444		
Observations	591		
Marginal R^2 / Conditional R^2	0.009 / 0.456		

Table 2.3. Model regressing predicted dissatisfaction scores on evaluation mode (joint vs. separate), outcome cause (self vs. other), their interaction, and subjects as random effects in study 4B.

	Dissat		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	6.23	6.09 – 6.38	**<0.001**
Agent [Self]	-0.73	-0.91 – -0.55	**<0.001**
EvaluationMode [S]	0.04	-0.16 – 0.24	0.700
Agent [Self] * EvaluationMode [S]	0.51	0.24 – 0.78	**<0.001**
Random Effects			
σ^2	1.42		
$\tau_{00\ ID}$	0.47		
ICC	0.25		
$N_{\ ID}$	1049		
Observations	1399		
Marginal R^2 / Conditional R^2	0.048 / 0.287		

REFERENCES

Alicke, M. D., & Govorun, O. (2005). The Better-Than-Average Effect. In Alicke, M. D., Dunning, D. A., & Krueger, J. I. (eds.). *The Self in Social Judgment. Studies in Self and Identity*. Psychology Press–106.

Buechel, E. C., Zhang, J., Morewedge, C. K., & Vosgerau, J. (2014). More intense experiences, less intense forecasts: Why people overweight probability specifications in affective forecasts. *Journal of Personality and Social Psychology*, *106*(1), 20–36.

Crisp, B. R., & Barber, J. G. (1995). The effect of locus of control on the association between risk perception and sexual risk-taking. *Personality and Individual Differences*, *19*(6), 841–845.

De Bruin, W. B., & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, *92*, 91–101.

DeCharms, R. (1968). *Personal causation: The internal affective determinants of behavior*. New York: Academic Press.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268.

Deci, E. L., & Ryan, R. M. (2013). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media

Frith, C. D. (2014). Action, agency and responsibility. *Neuropsychologia*, *55*, 137–142.

Gilbert, D. T., Morewedge, C. K., Risen, J. L., & Wilson, T. D. (2004a). Looking forward to looking backward: The misprediction of regret. *Psychological Science*, *15*, 346–350.

Gilbert, D. T., Lieberman, M. D., Morewedge, C. K., & Wilson, T. D. (2004b). The peculiar longevity of things not so bad. *Psychological Science*, *15*(1), 14–19.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*, 6.

Gneezy, A., Imas A., & Jaroszewicz, A. (2020). The impact of agency on time and risk preferences. *Nature Communications*, *11*, 1–9.

Heider, F. (1958). *The Psychology of Interpersonal Relations*, New York: John Wiley and Sons.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–157.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590.

Kahneman, D., & Miller, D. T. (1986). Norm Theory: Comparing Reality to Its Alternatives. *Psychological Review*, *93*(2), 136–153.

Kelley, H. H. (1967). Attribution Theory in Social Psychology. In Levine, D. (ed.) *Nebraska Symposium on Motivation*, (15), University of Nebraska Press, 192–238.

Kelley, H. H. (1973). The Processes of causal attribution. *American Psychologist*, *28*(2), 107–128.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328.

Larson, J. R. (1977). Evidence for a self-serving bias in the attribution of causality. *Journal of Personality, 45*(3), 430–441.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction?. *Psychological Bulletin, 82*(2), 213.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*, 1–12.

Morewedge, C. K., Gilbert, D. T., & Wilson, T. D. (2005). The least likely of times: How remembering the past biases forecasts of the future. *Psychological Science*, *16*(8), 626–630.

Morewedge, C. K., Kassam, K. S., Hsee, C. K., & Caruso, E. M. (2009). Duration sensitivity depends on stimulus familiarity. *Journal of Experimental Psychology: General*, *138*(2), 177–186.

Roth, S., & Kubal, L. (1975). Effects of noncontingent reinforcement on tasks of differing importance: Facilitation and learned helplessness," *Journal of Personality and Social Psychology*, *32*(4), 680–691.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28.

Tversky, A. (1972). Elimination by aspect: A theory of choice. *Psychological Review*, *79*, 281–299.

Vosgerau, J. (2010). How prevalent is wishful thinking? Misattribution of arousal causes optimism and pessimism in subjective probabilities. *Journal of Experimental Psychology: General*, *139*(1), 32–48.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (35, pp. 345–411). New York: Elsevier.

Wilson, T. D., & Gilbert, D. T. (2013). Comment: The impact bias is alive and well. *Journal of Personality and Social Psychology*, 105(5), 740–748.

Wortman, C. B. (1975). Some determinants of perceived control. *Journal of Personality and Social Psychology*, *31*(2), 282–294.

Chapter 3.

"DON'T FORGET THEM" OR "DON'T OVERLOOK THEM"? HOW THE NON-REVERSIBILITY OF A WORD IMPROVES MESSAGE EFFICACY

Giulia Maimone, Uma R. Karmarkar, On Amir

Rady School of Management, University of California, San Diego, La Jolla, CA, 92093, USA

ABSTRACT

Marketers have known for years that each word in a message can make a world of a difference, but less is known about the how and the why. Across five preregistered field and lab experiments (N = 22,024), we demonstrate when and how using an easily reversible (i.e., bi-polar) word in a statement, rather than a non-reversible one with the same meaning, engages different cognitive processes and leads to different outcomes. In particular, when a statement containing a bi-polar word is processed as a negation (i.e., opposing a claim rather than affirming it), a slower more elaborate cognitive process occurs. We show that this results in lower judgment confidence, and a lower likelihood to act on the message. In addition, we find that this more elaborative process also leads to weaker attitudes towards the message source. Our findings advance consumer theories by shedding light on the ways in which linguistic elements of communication impact judgments and real-world behaviors. They additionally offer practical persuasive messaging strategies for those engaged in a range of marketing and policy communications.

The consumer media experience is filled with persuasive messaging, from advertising products to political communications. As a consequence, when consumers encounter an assertion from a firm or individual, they immediately evaluate whether they find that statement to be true or false. However, these veracity judgments are not strictly binary. Instead, they are held with varying degrees of confidence, which in turn affects how likely people are to be persuaded and to act on the information in the future (e.g., Fazio and Zanna, 1978). These judgments can also relate to consumers' opinion of the message source, which has additional implications for future interactions with that entity. Past research has shown how message content and source characteristics can be central elements in shaping attitudes and persuasion (e.g., Karmarkar and Tormala, 2010; Petty and Cacioppo, 1986; Petty and Wegener, 1998; Priester and Petty, 1995; Tormala and Petty, 2004a). In this paper, we focus on the role of lexical choices made in crafting a message. Specifically, we show how the types of words used in a statement interact with consumers' judgment about the truthfulness of that statement, activating one of two distinct cognitive processes, and impacting downstream judgments, attitudes, and behaviors.

Imagine, for example, a brand communicating with the people of a town impacted by a hurricane using one of the two following statements:

Statement 1: "We will not overlook you."

Statement 2: "We will not forget you."

While these statements express the same sentiment, we argue they may lead to differential impacts, as "overlook" and "forget" are qualitatively different in terms of *word polarity*. "Forget" is bi-polar, meaning that it is easy to find a word that expresses its opposite (i.e., "remember"). In contrast, "overlook" is uni-polar, as finding a word that expresses its opposite is hard, and people tend to negate it by adding the tag "not" to the original concept (i.e., "not

overlook"). We find that, if consumers believe these statements to be true, the uni-polar statement (i.e., "We will not overlook you") engages a less elaborate and faster cognitive process that leads the message recipients to engage more with the message, experience higher levels of judgment confidence, hold stronger attitudes toward the brand and ultimately be more likely to take action.

THEORETICAL BACKGROUND

Given the importance of crafting effective consumer messaging in a crowded communication landscape, it is perhaps not surprising that several streams of research have used linguistics as a lens with which to study a variety of marketing relevant outcomes. For example, linguistic cues have been found to affect the efficacy of word-of-mouth communication as well as purchase intent (Packard and Berger, 2017). In designing a message, calibrating the choice of pronouns to the context improves customer satisfaction and product success (McFerran, Moore and Packard, 2019; Packard and Berger, 2020a; Packard, Moore and McFerran, 2018), while semantic concreteness boosts satisfaction with customer service, purchase intentions, and actual expenditures (Packard and Berger, 2020b). In addition, recent work has illustrated how linguistic and paralinguistic cues affect our perceptions of the message source. For example, using politically incorrect language was found to increase perception of source authenticity (Rosenblum, Schroeder and Gino, 2019). Building on these approaches, the current research helps understand and predict how and why certain linguistic elements can affect marketing-relevant judgments.

JUDGMENT CONFIDENCE AND ATTITUDE CERTAINTY

Confidence has been defined as the subjective sense of conviction about one's attitude or opinion (Festinger, 1950, 1954). Similarly, *judgment confidence* refers to a sense of conviction about one's judgments. Notably, high confidence in one's own judgments and thoughts about a message can increase the extent to which that message is persuasive (Petty, Briñol and Tormala, 2002). As a conceptually similar construct to judgment confidence, confidence in an attitude—or attitude certainty—is the metacognitive counterpart of attitude strength. High levels of attitude certainty increase an attitude's impact on a number of dimensions, from resistance to attitude change (e.g., Bassili, 1996; Tormala and Petty, 2002, 2004a), to the likelihood to translate attitudes into behavior (e.g., Fazio and Zanna, 1978; Fishbein and Ajzen, 1975; Gross, Holtz and Miller, 1995; Tormala and Rucker, 2007). For these reasons, these two constructs are particularly relevant in marketing contexts, where highly confident judgments and strong attitudes might make the difference between selling a product or not.

WORD POLARITY AND COGNITIVE PROCESSING

The reversibility of a concept expressed in a statement is captured by the term "word polarity" (Mayo et al., 2004). As previously described, a concept is bi-polar if it is relatively easy for people to retrieve its antonym. A concept is uni-polar if people find it challenging to retrieve its antonym (even if it exists) and describe its opposite via negating the original concept instead. Negations of uni-polar concepts engage a cognitive process described as *Schema-Plus-Tag (SPT*; Mayo et al. 2004; consistent with Gilbert, 1991; Grant, Malaviya and Sternthal, 2004). For example, when reading "We will not overlook you" in our earlier example, people process the affirmative uni-polar concept "We will overlook you" first (Schema), and then process its

negation by adding 'not' (Tag). In contrast, negations of bi-polar concepts engage a *Fusion* process. Thus, reading "We will not forget you" triggers antonym retrieval, which is directly substituted in, causing the statement to be processed cohesively as "We will remember you".[22] In this paper, we predict that Fusion (i.e., this retrieval-and-replacement process) is more effortful than Schema-Plus-Tag, and that this has important downstream consequences for consumer judgments.

THEORETICAL FRAMEWORK

In this work, we propose a framework that relates *word polarity* to several judgment and behavioral outcomes through distinguishable cognitive processes. To do so, we first consider the *claim type* of a statement, meaning whether it is an affirmation (e.g., "We will overlook") or a negation (e.g., "We will not overlook"). We then introduce a novel factor, distinct from the cognitive linguistics literature, based on the fact that in real life people are not merely passive listeners. Consumers often have relevant priors and beliefs and use them to judge whether these messages are truthful (e.g., whether they think the brand will, or not, overlook the impacted community). Our research explores how the inclusion of these judgments can change the way a statement is encoded (i.e., framed in the consumer's mind). We thus define *statement encoding* as the interaction of a statement's *claim type*, and people's *judgment of the truthfulness* of that statement (i.e., affirmation vs. negation x true vs. false; Figure 3.1). This allows us to identify three conceptual classes illustrated in Figure 3.1: 1) *Affirmation encoding:* judging an affirmation

[22] Note that affirmations are processed as SPT regardless of their polarity, because, in absence of the tag 'not', people process only the Schema. Similarly, double negations are also processed as SPT regardless of their polarity, because, by definition, in all double negations the tag 'not' is eliminated, leaving only the Schema to be processed.

claim type as 'true', 2) *Negation encoding*: judging a negation claim type as 'true' or judging an affirmation claim type as 'false', and 3) *Double Negation encoding:* judging a negation claim type as 'false'. So, if a consumer believes the brand's statement "We will not overlook you" to be true, that combines a negation claim type with a 'true' judgment, leading to a *Negation encoding*. If, a consumer believes that statement to be false, we have a negation claim type and a 'false' judgment, leading to a *Double Negation encoding*.

Using this novel approach, we propose that only *Negation encodings* of statements employing *bi-polar* words engage the Fusion process, while all the other combinations of statement encoding and word polarity engage Schema-Plus-Tag (see Figure 3.2). Thus, we can predict that the bi-polar negation statement "We will not forget you" would engage different cognitive processes dependent on whether a listener judged it to be true or false.

HYPOTHESES

The primary distinction between the Fusion and Schema-Plus-Tag processes is that Fusion requires the additional cognitive function of actively retrieving a word's antonym. Thus, Fusion involves more elaborate—and likely more effortful—processing than SPT. This leads us to three hypotheses. We predict that the more effortful Fusion process will lead to lower message efficacy (**H1**). In addition, we hypothesize that this would be the case because the Fusion process will lead listeners to have lower confidence in their judgments about the truthfulness of a statement (**H2**). Finally, given the conceptual similarity between judgment confidence and attitude certainty, we predict that the statements engaging the Fusion process will result in weaker post-message attitudes (**H3**).

To better understand the implications of these hypotheses, consider again the seemingly

equivalent Statement 1 (“We will not overlook you”) and Statement 2 (“We will not forget you”) from the brand messaging example. Assuming a consumer believes the brand’s intent, both of these statements are *Negation encodings*. However, as mentioned, “overlook” in Statement 1 is uni-polar, leading to the SPT process, while “forget” in Statement 2 is bi-polar, engaging the Fusion process (see Figure 3.2). As a result, we predict that Statement 2 (forget) would generate lower engagement with the message, will make consumers feel less confident in their judgments of truthfulness of the statement, and will result in weaker attitudes than Statement 1(overlook).

LINGUISTIC PROCESSING INFLUENCE ON ATTITUDE STRENGTH: MESSAGE OR MESSENGER?

Numerous prior studies have investigated the impact of characteristics of a message’s source on message interpretation and persuasion, both in positive and negative directions (e.g., Priester and Petty, 1995; Petty and Wegener, 1998; Tormala and Petty, 2004b). Substantially less attention has been devoted to a potential opposite causal pattern, i.e., the impact that the characteristics of a persuasive communication can have on attitudes toward the messenger.

While we predict stronger attitudes arising from SPT, differences arising from how the message is cognitively processed could relate to perceptions of the message subject, the message source, or both. *Statement encoding* is intimately related to the content of the message, but the message itself also represents the voice of the speaker. In our example, the brand is sending a message about itself, therefore the listener’s judgments and attitudes relate to both the message subject and the message source. However, a single firm might communicate about multiple brands. Similarly, an independent influencer or reviewer might choose to discuss more than one product. As a result, it is useful to understand whether the predicted effects of message wording are

influencing attitudes towards the message subject, the message source or both, and this is investigated in Study 4.

STUDIES OVERVIEW

We test our hypotheses in five preregistered experiments. Study 1 examines whether marketers can use word polarity in a message to maximize the efficacy of their communications. Partnering with a non-profit organization, we provide field evidence supporting H1, showing that the choice of one word (uni- vs. bi-polar) generates different Facebook click through rates. Given this demonstration of meaningful and consequential impact, the subsequent experiments offer detailed support of the underlying theoretical framework. Study 2 confirms how the interaction of *statement encoding* and *word polarity* causes, indeed, two different cognitive processes via the use of response time measures and how these processes influence judgment confidence. As predicted by H2, we find that statements engaging the Fusion process generate lower judgment confidence.

Building on this, Study 3A employs a persuasive political message context and finds that language engaging the Fusion process leads to less extreme attitudes than the same content using language engaging SPT, in line with H3. Study 3B replicates these results with additional controls for potential differences in perceived meaning between the uni-polar and bi-polar statements. Study 4 extends H3 by disentangling whether our observed effects are impacting the recipient's attitude toward the message subject, the message source, or both, in a consumer context involving a product review. This experiment demonstrates that encoding differences arising from lexical choice have more impact on perceptions of the message source (i.e., the

reviewer) than the message subject (i.e., the product). Collectively, our findings show that the reversibility of a statement has a significant impact on a number of dimensions ranging from judgment confidence to attitude extremity to marketplace behavior.

STUDY 1

We predict that a message engaging the Fusion process will be less effective at eliciting behavior than one engaging SPT (H1). To demonstrate the consequential importance of this effect, Study 1 tests this hypothesis in a field experiment using the crowded attention marketplace of Facebook ads.

We partnered with the non-profit organization Project Kesher. During the 2022 military conflict in Ukraine, Project Kesher has been working to support women in that region with emergency cash grants, by establishing a charitable fund, and by overseeing the creation of emergency kits for feminine care. In this study, we compared click through rates on Project Kesher Facebook advertisements that used messages designed to engage either Fusion or SPT processing.

METHODS

The stimuli for this study were defined by a pre-test used to identify uni-polar/bi-polar pairs of messages with similar meanings. Based on the results, the pair of statements “Do not overlook them” (uni-polar) and “Do not forget them” (bi-polar) were chosen. See Web Appendix for the pre-test details.

Material and procedure. This study is preregistered on AsPredicted (https://aspredicted.org/56K_38F). As a field experiment, the study is designed to compare the performance of two advertisements using Facebook's split testing platform, which offers an A/B test function that employs random assignment (Orazi and Johnston, 2020). Project Kesher ran the A/B test from their Facebook manager account, specifying the campaign objective as "traffic" (i.e., clicks to visit the Project Kesher Facebook page linked to the advertisement). The target audience was defined as users who live in the United States, older than 18 years old, and with similar characteristics to those who liked Project Kesher's page.[23] As preregistered, the sample size was determined by a daily budget limit of $100 for five days. At the end of this time period, we reached a total of 20,118 unique Facebook users, and generated 26,379 impressions (i.e., total advertisements visualizations).

Both advertisements featured an image depicting a group of women and children holding hand-written signs and a Ukrainian flag, with the headline "Support Ukrainian women and girls". Above the image was either the text "Do not forget them" (Bi-polar statement), or "Do not overlook them" (Uni-polar statement; see Figure 3.3). Because the messages are both negation claim types and, as imperatives, they can only elicit a "true" judgment, they are both *Negation encodings.* According to our framework, the bi-polar message would engage the Fusion process, and the uni-polar one the SPT process. For this reason, we will refer to the bi-polar message "Do not forget them" as the 'Fusion condition', and to the uni-polar message "Do not overlook them" as the 'SPT condition'.

[23] More specific parameters or the algorithm to define "similar to users who liked Project Kesher's page" were not disclosed.

RESULTS

As preregistered, the dependent variable of interest was the ratio between the total number of clicks the advertisements received, or "total link clicks",[24] and the impressions generated in each condition.[25] The advertisement with the message "Do not overlook them" (i.e., the SPT condition) generated 11,084 impressions, and was clicked 525 times, meaning it obtained a 4.74% click through rate. The advertisement with the message "Do not forget them" (i.e., the Fusion condition) generated 15,295 impressions, and was clicked 623 times, yielding a 4.07% click through rate. A logistic model regressing total link clicks on cognitive process conditions, using total impressions as the sample (N = 26,379), found that these click through rates were significantly different from each other ($b = 0.16$, $z = 2.60$, $p = .009$).

This finding supports our hypothesis that the advertisement with the message engaging the Schema-Plus-Tag process would be more effective than the one with the message engaging the Fusion process. Finally, the difference in click through rates between the two advertisements has financial implications. Facebook uses the total amount of money spent on each advertisement and the total link click metric to define the cost-per-click of each advertisement. Based on this, the cost-per-click of the advertisement engaging the SPT process ("overlook"; $0.20) was significantly lower than that of the advertisement engaging the Fusion process ("forget"; $0.23).

Despite social media being a notoriously difficult channel for attracting consumer attention in general, these results still show a significant impact of word polarity, demonstrating that these lexical choices can be used strategically to improve advertising strategies and increase

[24] Note that for this metric, if a single user clicked on the link twice, this would be counted as two separate clicks.
[25] This analysis is preregistered as the secondary analysis. We could not perform the primary analysis, as Facebook did not provide Project Kesher with the Unique Link Clicks metric (i.e., the number of *unique* Facebook users who clicked on the advertisement). We have no theoretical reason to suspect such analysis would have generated a different result.

social media engagement. Taking a step back from these outcomes, our framework predicts that the mechanisms driving these effects arise from word polarity engaging different cognitive processes and impacting judgment confidence and attitudes. In the following studies, we provide evidence defining the building blocks of those mechanisms.

STUDY 2

As discussed, Study 1 showed the behavioral and market impact of word polarity in advertising in the field. Study 2 tests our underlying theoretical process framework providing evidence that word polarity interacts with statement encoding to elicit two distinct cognitive processes with different effects. Specifically, it tests our prediction that the Fusion process will lead to lower judgment confidence (H2). As illustrated across Figures 1 and 2, there are three factors contributing to the proposed effect. The first two are *word polarity* (uni-polar vs. bi-polar) and *claim type* (affirmation vs. negation). In this work we introduce the consumers' *veracity judgment* of the statement (true vs. false) as a third factor. Nominally, this would suggest a 2x2x2 design. However, since our focus is on comparing Fusion and SPT, we can combine these factors into two conditions defined by the two processes (see Figures 1 and 2).

METHODS

We ran two pre-tests to accurately select the stimuli for Study 2. In Pre-test 1, to define an appropriate list of bi-polar and uni-polar words, participants were asked to generate antonyms for a range of words and rate the difficulty of doing so (see Table 3.1). Pre-test 2 was designed to identify stimuli for which participants were likely to hold a pre-existing true-or-false belief.

Using these results, we created two types of stimuli: those for Statement-type trials (***S***) and those for Question-type trials (***Q***; Table 3.2). The Question-type trials (e.g., "Do you have very basic knowledge in American history?") were used to define exclusions, they were chosen to have answers that were likely to predict whether participants had pre-existing beliefs about the related Statements-type trials (e.g., "John F. Kennedy was elected before Ronald Regan"). Details regarding both pre-tests are available in the Web Appendix.

Material and procedure. This study is preregistered on AsPredicted (https://aspredicted.org/KGA_LOL). Participants were recruited through a behavioral laboratory at a large west coast university. While demographic information was not collected, the subject pool is drawn from the university's undergraduate population, and participation in the lab is not restricted by any demographic variables. Sample size was pre-set as one week of data collection in the lab. Two hundred and seventy undergraduates participated and were compensated with course credit. To incentivize their performance on both accuracy and speed, participants learned that an algorithm would identify the five participants who correctly judged the veracity of the greatest number of *statements* with the best-calibrated confidence ratings in the shortest time (as measured by fastest response times). These five people were awarded a $10 Amazon.com gift certificate.

Each participant made multiple judgments, resulting in a repeated-measures design. The experiment was conducted using DirectRT software (Empirisoft), allowing for measurement of behavioral responses and response time (RT) in milliseconds. The full design, including participant instructions, is available in the Web Appendix. All participants viewed instructions explaining that they were going to see some Question-type trials, and that they had to answer

each one by either pressing the key "Y"—for 'yes'—or the key "N"—for 'no'. During these ratings, each Question-type trial was shown on a different screen.

After answering the 11 Question-type trials defined via Pre-test 2 (see Table 3.2), participants were instructed that the following screens would show different Statement-type trials, presented one at a time. They were asked to judge each of them by either pressing the key "T"—if they thought the Statement-type trial was true—or the key "F"—if they thought the Statement-type trial was false. Response time was measured based on the time from the appearance of the Statement-type trial on the screen to when the participant pressed an answer key. Participants were also asked to rate their confidence in their judgment on a scale from 1 [completely uncertain about my answer] to 9 [completely certain about my answer] using the keyboard. All participants judged two practice Statement-type trials (not included in the analysis), and then the 12 pre-tested Statement-type trials (see Table 3.2), presented in randomized order.

RESULTS

As mentioned above, our stimuli reflected Question-type trials chosen to be predictive of whether or not participants had pre-existing beliefs about the related Statement-type trials. Thus, as preregistered, if participants answered "no" to a Question-type trial, we dropped judgment and confidence observations for any related Statement-type trial. These exclusions left 2,093 observations total across 268 participants (i.e., 2 participants answered "no" to all Question-type trials) with an average of 4.25 observations (SD = 2.56) excluded per participant.

As described, the Fusion process is more complex and effortful than SPT, so we expected statements engaging Fusion would take longer for participants to process. Thus, as a manipulation check, we ran a linear mixed effects model regressing judgment RT (i.e., response

time) on process type, with random effects for subjects and Statement-type trials. This model confirmed that Statement-type trials leading to Fusion resulted in longer RTs than statements leading to SPT (b = 655.56, t(1,619.36) = 2.63, p = .009). This finding supports our theoretical framework proposing that claim type, judgment of veracity, and word polarity interact to elicit different cognitive processes, and, more importantly, confirmed that our manipulations tapped such different cognitive processes.

The main question we intended to test in this study was whether statements engaging the Fusion process would lead to lower judgment confidence than those engaging SPT (H2). We ran a linear mixed effects model regressing judgment confidence on process type, with random effects for subjects and Statement-type trials.[26] As predicted by H2, the effect of process type on confidence was significant such that Statement-type trials engaging the Fusion process generated lower confidence in one's own judgment than those engaging SPT (b = -0.49, t(1,715.02) = -3.54, p < .001). When consumers encounter claims in the marketplace, they are likely to hold opinions about whether these claims are true or false. These results provide strong evidence linking the cognitive processes evoked by negations with perceptions of statement truth to judgment confidence.

It is unlikely that the confidence effect was driven by specific stimuli (i.e., Statement-type trials) used, as opposed to the cognitive processes they engaged. Since SPT/Fusion conditions are partially determined by participants' true/false judgments of each Statement-type trial (Figure 3.1), some Statement-type trials in this experiment appeared in both conditions, depending on

[26] The random effects model used here is slightly different from the pre-registered one. The model design specified in the pre-registration yields the same qualitative effect at a higher level of significance (b = 1,066.0, t(1,435.30) = 4.63, p < .001). We judged the model in the main text as more appropriate, because the Statement-type trials' order was already randomized by design.

whether a particular participant judged them to be true. Moreover, the confidence effect observed appears to be independent from other word characteristics like length. For example, the bi-polar words identified in Pre-test 1 happen to be shorter, and more common than the uni-polar words. This makes our experiment a conservative test of our hypothesis, because the statements engaging Fusion took longer to process and led to lower judgment confidence, despite Fusion only being engaged by the shorter and more common bi-polar words.

STUDY 3A

Study 2 offered a rigorous test of our psychological framework. It showed evidence supporting our proposition that word polarity interacts with statement encoding engaging two types of processes (e.g., RT differences) and showed systematic differences in judgment confidence arising from those distinct cognitive processes. In the following studies, we expand the scope of the implications of this difference across multiple forms of persuasive messaging consumers might encounter. Given the conceptual similarity between judgment confidence and attitude certainty, in Studies 3A and 3B, we hypothesize that the lower judgment confidence level generated by the Fusion (vs. Schema-Plus-Tag) process will also translate to less extreme attitudes toward the message (H3).

METHODS: MATERIAL AND PROCEDURE

This study is preregistered on AsPredicted (https://aspredicted.org/WLY_ZQP). Participants were recruited on Amazon Mechanical Turk and engaged in the study for monetary compensation.

Participants were randomly assigned to one of two conditions. They read a scenario describing a political candidate running for office at the city level who had answered a question from the public by saying either "During my term in office, I will not accept any forms of bribery" (bi-polar/reversible statement), or "During my term in office, I will not tolerate any forms of bribery" (uni-polar/non-reversible statement). They then indicated whether they believed the candidate's statement to be true or false. Based on their answer, participants were asked a valenced-matched question designed to elicit their attitude towards the candidate. Those who rated the candidate's statement about refusing bribes as true were asked "To what extent do you agree with the following statement: 'This candidate is an ethical person'". Those who rated the candidate's statement as false were asked "To what extent do you agree with the following statement: 'This candidate is an unethical person'". Ratings took place on a scale from 0 (Neither disagree nor agree) to 6 (Strongly agree). Note that we operationalize attitude strength in terms of the specific marker of attitude extremity (Abelson, 1995; Tormala and Rucker, 2022). As a manipulation check, depending on condition, participants were then asked to write an antonym for the word 'accept' or 'tolerate,' and completed the experiment by answering an attention check and demographic questions.

As discussed in the introduction, our theoretical framework allows us to combine the three different statement elements (*claim type*, *word polarity* and *judgment of veracity*) into one comparison (SPT vs. Fusion; see Figures 1 and 2). Both statements (i.e., "I will not accept[tolerate] any forms of bribery") are negation *claim types*. As per our theory, only recipients who believe the bi-polar statement is true (bi-polar *Negation encoding*) will engage the Fusion process. Participants will engage the SPT process if they believe a) the bi-polar statement is false (bi-polar *Double Negation encoding*), b) the uni-polar statement is false (uni-polar

Double Negation encoding), or c) the uni-polar statement is true (uni-polar *Negation encoding*). Expressions of attitude extremity could differ between positive and negative attitudes. Thus, the most appropriate comparisons would be between true judgments or between false judgments. However, since false judgments of the uni-polar and bi-polar statements both lead to SPT, we focused—as preregistered—on the comparison between participants who believed the uni-polar statement was true (Schema-Plus-Tag condition), versus those who believed the bi-polar statement was true (Fusion condition). This necessitated exclusion of all participants who indicated that they believed the statements to be false.

Given the potential scope of the preregistered exclusions, the sample size was pre-set to 400 participants.[27] Participants who failed an open-ended attention check question were also excluded. Four hundred and one participants completed the experiment and 29 failed the attention check. Before excluding participants who judged the statements to be false, we ran two manipulation checks. One to test whether the words 'accept' and 'tolerate' were perceived as bi-polar (reversible via an antonym) and uni-polar (non-reversible), respectively. As predicted, more participants could find an antonym for 'accept' (82.26%) than for 'tolerate' (3.23%; χ^2 (1) = 135.91, $p < .001$, N = 372), classifying them as intended. The second manipulation check was used to confirm whether the two statements were judged similarly in terms of veracity. As expected, the same proportion of participants judged the candidate's statements as true in the uni-polar (75.8%) and bi-polar conditions (73.7%; χ^2 (1) = 0.13, $p = .720$, N = 372).

Following these tests, we excluded 94 participants who believed the candidate's statement to be false. Our sample for the main analysis therefore consisted of 278 observations (44.96% female,

[27] We preregistered collection of an additional sample of 200 participants if the number of individuals surviving exclusion per condition was lower than 50 in the primary data collection. However, this was not needed, and no additional recruitment or data collection took place.

M_{Age} = 37.69, SD_{Age} = 12.28). Additional details, including participant instructions, are available in the Web Appendix.

RESULTS

As preregistered, we tested whether the Fusion condition ("I will not accept") led to weaker attitudes towards the message than the SPT condition ("I will not tolerate"; H3). We regressed attitude extremity on process type. As predicted, we found a main effect of process type such that engaging the Fusion process generates a less extreme (less positive in this case) attitude toward the candidate's ethical stance compared to engaging the SPT process (b = -0.39, $t(276)$ = -2.94, p = .004). In line with H3, this experiment demonstrates that our framework can predict how differences in word choice change attitude extremity in persuasive messages.

STUDY 3B

One challenge with using different words between the SPT and Fusion conditions is whether the attitude extremity effect might have arisen from any semantic differences. To provide additional evidence that our findings in Study 3A are caused by differences in cognitive processing rather than such differences in semantic interpretation, Study 3B replicates the effect while controlling for the inferred meanings of the words.

METHODS: MATERIAL AND PROCEDURE

This study is preregistered on AsPredicted (https://aspredicted.org/XFZ_YJQ). Participants were recruited on Amazon Mechanical Turk and engaged in the study for monetary

compensation.

Materials and procedures were the same as in Study 3A, except for the addition of the following open-ended question: "What do you think 'During my term in office, I will not tolerate[accept] any form of bribery' means?" As preregistered, these "meaning" responses were coded into the following pre-set categories: 1 = "I will not take bribes", 2 = "I will not take bribes and will punish those who do", 3 = "other", 4 = general negative comment (e.g., "politicians are all liars"), or 5 = irrelevant (e.g., "I like ice-cream"). Coding was done blind to experimental conditions.

As in Study 3A, and as specified in the preregistration, participants who failed an attention-check question and those who judged the candidate's words to be false were excluded from analysis. In addition, we pre-committed to exclude participants whose "meaning" answer was coded as either 4 (general negative statement) or 5 (irrelevant). Given that these parameters had the potential to exclude a significant number of the participants, and we intended to match the sample size of Study 3A (i.e., N = 278), the desired sample size was set to 600.[28] Five hundred and ninety-eight participants completed the experiment, with 26 failing the attention check, and 178 being excluded based on their "meaning" responses. Replicating the manipulation checks of Study 3A, regardless of whether participants had judged the statement as true or false, they were significantly more able to generate an antonym for 'accept' (86.50%) than for 'tolerate' (0.52%; χ^2 (1) = 170.02, $p < .001$; N = 394), indicating that they were perceived as more bi-polar and uni-polar, respectively. Additionally, the same proportion of

[28] We pre-registered collection of an additional sample of 300 participants if the number of individuals surviving exclusion per condition was lower than 100 in the primary data collection. However, this was not needed, and no additional recruitment or data collection took place.

participants judged the candidate's statements as true in the uni-polar (79.4%) and the bi-polar conditions (79.0%; χ^2 (1) < 0.001, p = 1; N = 394).

Finally, 82 participants in the remaining pool were excluded based on believing the candidate's statement to be false. Our sample for the main analysis therefore consisted of 312 observations (50.64% female, 0.96% other, M_{Age} = 39.53, SD_{Age} = 12.08). Additional details, including participant instructions, are available in the Web Appendix.

RESULTS

As in Study 3A we regressed attitude extremity on process type for participants who judged the political candidate's statements as true. As predicted, we replicated the main effect of process types on attitude extremity, such that engaging the Fusion process generates a less extreme attitude toward the candidate compared to engaging the SPT process (b = -0.32, t(310) = -2.50, p = .013).

As preregistered, we conducted a second version of this analysis controlling for potential variation in interpreted meaning. Comparing the relative coded meaning perceptions (categories 1, 2 and 3), we found differences between conditions (χ^2 (2) = 37.22, p < .001). Specifically, "I will not take bribes" (category 1) was reported more often for the statement including the word "accept" (81.01%) than it was for the statement with "tolerate" (55.84%; χ^2 (1) = 8.24, p = .004). In contrast, "I will not take bribes and I will punish those who do" (category 2) was reported more often for "tolerate" (22.08%) than for "accept" (1.27%; χ^2 (1) = 28.44, p < .001). Responses falling in the "other meanings" (category 3) were similar across conditions (χ^2 (1) = 0.58, p = .446). Regressing attitude extremity on process type together with the 'Coded Meaning' variable (answer variable: "1" = 0, "2" = 1, and "3" = 2) still replicated the significant effect of lexical

choice. The Fusion condition generated less extreme attitudes toward the candidate than the SPT condition (b = -0.31, $t(308)$ = -2.29, p = .023). This provides convergent evidence that the impact of lexical choice arose from the predicted differences in cognitive processing above and beyond semantic distinctions. Collectively, Studies 3A and 3B find that the polarity of word choice can affect consumer judgments in ways that impact attitude extremity in the domain of persuasion.

STUDY 4

Study 4 examines how word polarity can influence attitudes in the context of product reviews. In Studies 3A and 3B we found that engaging the Fusion process leads to less extreme attitudes in a scenario where a politician was making an assertion about themselves. An additional question arising from this is whether the effect of differing cognitive processes was impacting message recipients' attitudes toward the message subject, the message source, or both. To explore this, we examined source and subject perceptions separately.

METHODS

The stimuli for this study were defined by two pre-tests. Pre-test 1 was used to identify bi-polar/uni-polar pairs of words with similar meanings. Two pairs of words emerged from Pre-test 1 (i.e., uninteresting and monotonous, and unsatisfactory and lame). Pre-test 2 tested the 'reversibility' of four statements containing these words and demonstrated that the statements containing the "uninteresting-monotonous" pair had the most clearly defined polarities (see Web Appendix for details).

Material and procedure. This study is preregistered on AsPredicted (https://aspredicted.org/EPM_ANY). Participants were recruited on Amazon Mechanical Turk and engaged in the study for monetary compensation.

Participants read a scenario describing a reviewer writing about the first novel from an up-and-coming writer. The reviewer seemed to like the book, and during the review they noted either "This book is not uninteresting" (Bi-polar statement; Fusion condition), or "This book is not monotonous" (Uni-polar statement; SPT condition). We employed this design on a sample separated into two pools with distinct dependent variables. For the first pool, we randomly assigned participants to either the Fusion or the SPT conditions and asked them to indicate their attitude towards the message **source**. For the second pool, we randomly assigned participants to either the Fusion or SPT conditions (using the same stimuli), and asked them to indicate their attitudes towards the message **subject**.

As in Studies 3A and 3B, both statements about the book are negation *claim types*, and we preregistered to exclude all participants who judged the statements as false. Using Studies 3A and 3B as a reference, we estimated a false judgment rate of about 15% and committed to collect 1,200 observations.[29] As pre-registered, participants who failed the open-ended attention check question were also excluded. One thousand two hundred and five participants completed the experiment, 25 failed the attention check, and 132 believed the candidate's statement to be false. Our final sample included 1,048 participants (50.86% female, 1.05% other, $M_{Age} = 38.19$, $SD_{Age} = 12.18$).

[29] Our pre-registration plan included the intent to collect an additional 400 observations in case we had less than 200 qualifying participants per condition after the false judgments exclusion. However, this was not necessary, and no additional data was collected.

All participants indicated whether they believed the reviewer's statement to be true or false. As noted, one pool of participants expressed their attitude toward the reviewer (message source) in the Fusion and SPT conditions, and the second expressed their attitude toward the book (message subject) in the Fusion and SPT conditions. Those in the message source group were asked "To what extent do you agree with the following statement: 'This reviewer is competent'". Those in the message subject group were asked "To what extent do you agree with the following statement: 'This is a good book'". Both questions were answered on a scale from 0 (Neither disagree nor agree) to 6 (Completely agree). Participants were then asked to answer the attention check and demographic questions.

RESULTS

Regressing attitude towards the message source on cognitive process type we find that participants in the Fusion (versus SPT) condition had less positive perceptions of the reviewer (b = -0.50, $t(518)$ = -4.14, $p < .001$). This effect replicates the findings of Studies 3A and 3B by showing that the differing cognitive processes evoked can influence attitude extremity. However, regressing attitude towards the message subject on process type showed no effect of Fusion on attitudes toward the book (b = -0.08, $t(526)$ = -0.83, p = .409; see Figure 3.4).

Attitudes toward the message subject and the message source were measured on the same scale (0-6) and were collected on distinct samples. This creates four groups defined by two levels associated with the two factors. Thus—as preregistered—we are able to conduct a two-way type 3 ANOVA to examine whether there is a difference-in-difference between the impact on attitudes towards the message source and message subject. Again, replicating Studies 3A and 3B, there was a significant main effect of Process Type ($F(1, 1{,}044)$ = 14.14, $p < .001$, η_p^2 = .013)

such that engaging the Fusion process led to less extreme (less positive; M = 3.85, SD = 1.30) attitudes than SPT (M = 4.14, SD = 1.23). Attitude Type also showed a main effect (F(1, 1,044) = 14.69, $p < .001$, $\eta_p^2 = .014$) such that the attitude toward the message subject (the book; M = 4.14, SD = 1.11) was more extreme than the attitude toward the message source (reviewer; M = 3.86, SD = 1.40). Beyond this, we observe a significant interaction (F(1, 1,044) = 7.44, $p = .006$, $\eta_p^2 = .007$) such that the effect of Process Type was stronger on attitude toward the reviewer (message source) than on attitude toward the book (message subject; see Figure 3.4). Thus, we demonstrate that encoding differences arising from lexical choices have more impact on perceptions of the message source than the message subject.

GENERAL DISCUSSION

We demonstrate how the type of words used in a message can interact with consumers' veracity judgments to engage different types of cognitive processes that affect the message's efficacy, consumers' judgment confidence in the validity of the message, and their attitude extremity toward the message source. Building on cognitive linguistic theory, we created a conceptual framework that considers the integration of *word polarity, linguistic claim type,* and consumers' *judgment of statement truthfulness*. This allows us to predict judgment-relevant outcomes via a parsimonious categorization metric (see also Figure 3.2).

The present studies explore predictions of our framework for consumer behavior in the field (response to advertisements), as well as attitude formation in settings like political speech and product reviews. We find that different process-dependent encodings can influence a message recipient's resultant likelihood to act on the message and attitude extremity. Generally

speaking, the purpose of persuasive messages is to influence behaviors and attitudes about the message topic. That is, a book review or other form of product advertisement is aimed at influencing the recipient's perception of the product. Interestingly, we find that our effect influences attitudes towards the message *source*, subsequently impacting consumers' likelihood to engage with the product (as observed in Study 1). Also, since these elements align during self-promotion (Studies 3A/B), our effect has potential benefits for firms in which the product and company brand are unified, such as Tesla or AirBnB.

Our findings offer consumers, marketers, politicians, and policy-makers practical guidelines for increasing the persuasiveness of their communications, and positive attitude strength towards themselves and/or their brand. Studies 3A and 3B showed how powerful such guidelines could be for policy and political messaging—certainly relevant during elections, when the public turns a critical eye to the onslaught of campaign claims, often drawing from prior judgments about the candidates or parties. Additional interesting domains of applications for these findings may include health care messaging, where compliance with health care recommendations are not only beneficials to individuals, but also to the communities. Finally, our results from the field and our findings about writing product reviews illustrate the usefulness of this insight related to designing advertising campaigns and corporate messaging more broadly.

From a theory perspective, we show how consumers' judgments of message veracity actively relate to findings in the negation processing literature in psychology (Gilbert, 1991; Gilbert et al., 1990; Mayo et al., 2004) thereby strengthening the conceptual link between cognition and feelings of confidence. Overall, it offers a novel cross-disciplinary framework built on cognitive linguistics that can be used to better understand attitudes, behavior and persuasion in policy and marketing domains.

ACKNOWLEDGEMENTS

The authors thank Victor Ferreira, Grant Packard, Thomas Urbach, S. Christian Wheeler, Rand R. Wilcox, and Piotr Winkielman for helpful comments, and are grateful to Samantha Tieger and Project Kesher for their collaboration.

Chapter 3, in full, has been submitted for publication of the material. Maimone, Giulia, Uma R. Karmarkar and On Amir. The dissertation author is the primary investigator and author of this paper.

DATA AND MATERIALS AVAILABILITY

All data, analyses code, and material are available at the following link: https://osf.io/kybhx/?view_only=b92c8b5f3e8c42ff8dbf8ef959174cec. Please contact Giulia Maimone at giulia.maimone@rady.ucsd.edu for any additional questions.

All studies' design and analysis were preregistered on AsPredicted (Study 1: https://aspredicted.org/56K_38F, Study 2: https://aspredicted.org/KGA_LOL, Study 3A: https://aspredicted.org/WLY_ZQP, Study 3B: https://aspredicted.org/XFZ_YJQ, Study 4: https://aspredicted.org/EPM_ANY).

FIGURES

		Claim Type	
		Affirmation	Negation
Truthfulness Judgement	True	*Affirmation encoding*	*Negation encoding*
	False	*Negation encoding*	*Double Negation encoding*

Figure 3.1. Statement Encoding Matrix

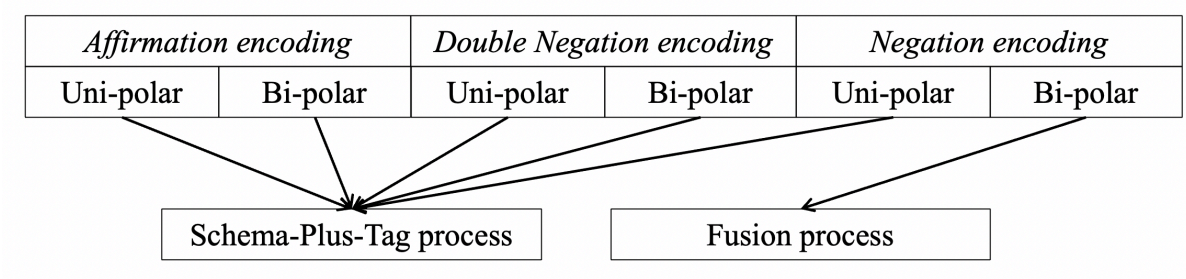


Figure 3.2. Theoretical framework

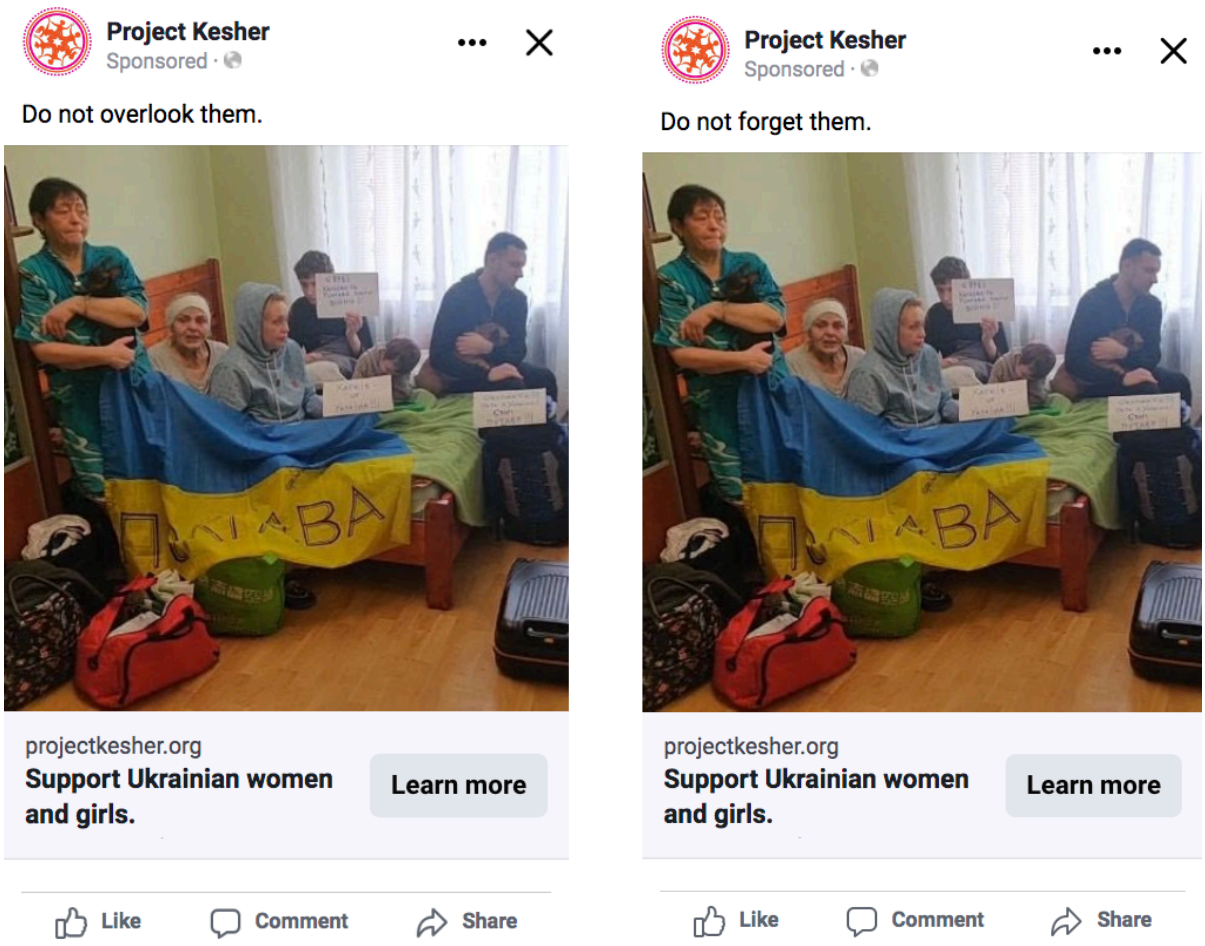


Figure 3.3. Facebook advertisements in the SPT (left panel) and in the Fusion (right panel) conditions in Study 1.

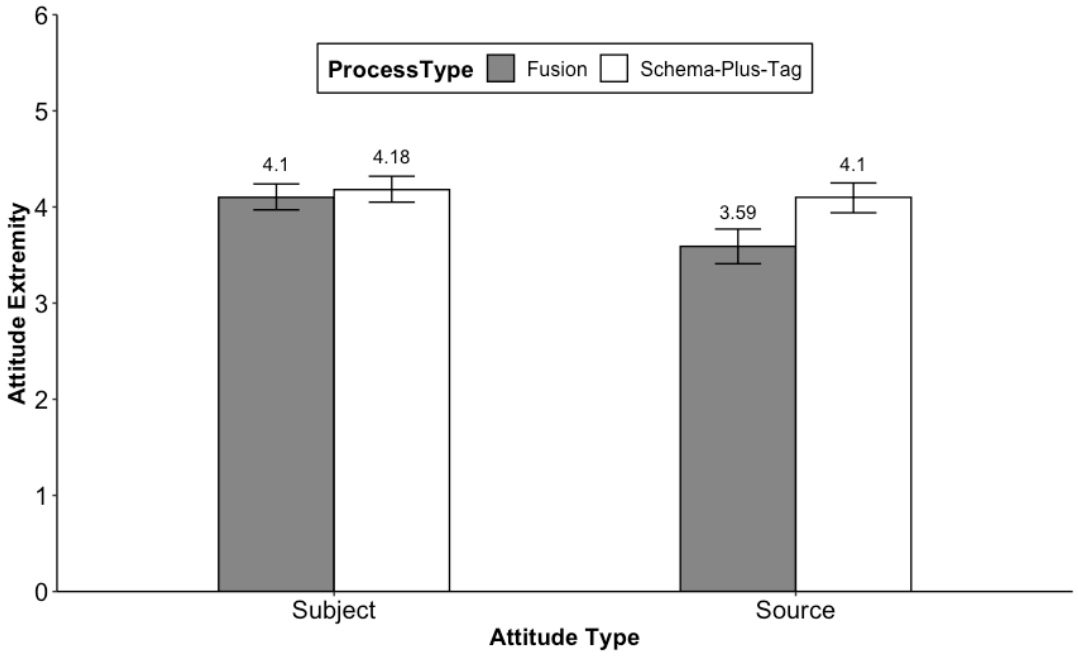


Figure 3.4. Effect of process type on attitudes toward the message source in Study 4. Error bars reflect 95% CIs.

TABLES

Table 3.1. Bi- and uni-polar words selected based on Pre-test 1 of Study 2

Bi-polar	**Uni-polar**
Before	Caloric
More	Deaf
Heavy	Precious
Negative	Territorial
Short	Charismatic
Northern	Insulating
Noisy	Eccentric
	Genetic

Table 3.2. Study 2 stimuli arising from Pre-test 2

	Affirmations	**Negations**
Bi-polar	***S****:* John F. Kennedy was elected before Ronald Regan ***Q***: Do you have very basic knowledge in American history?	***S***: Turtles can not live more than 100 years ***Q***: Do you have very basic general knowledge?
	S*:* The charge of electrons is negative ***Q***: Do you have very basic knowledge in chemistry?	***S****:* Italy is not in the Northern hemisphere ***Q***: Do you have very basic knowledge in geography?
	S*:* One liter of oil is heavier than one liter of water ***Q***: Do you have very basic knowledge in natural sciences?	***S****:* Explosions in outer space are not noisy ***Q***: Do you have very basic knowledge in acoustics?
Uni-polar	***S****:* Dogs are deaf when they are born ***Q***: Do you have very basic knowledge about animal development?	***S****:* Avocado is not a caloric fruit ***Q***: Do you have very basic knowledge about food properties?
	S*:* Vincent Van Gogh had an eccentric personality ***Q***: Do you have very basic knowledge in art history?	***S****:* Glass is not an insulating material ***Q***: Do you have very basic knowledge in material science?
	S: Down Syndrome is a genetic disorder ***Q***: Do you have very basic knowledge in genetics?	***S****:* Traumas are not genetic conditions ***Q****:* Do you have very basic knowledge in genetics?

DESCRIPTION OF THE PRE-TEST OF STUDY 1

We identified three potential messages for the advertisement, and created a uni-polar and a bi-polar version of each. Each message was approved by Project Kesher's marketing team. The current pre-test was used to select a pair of stimuli (i.e., statements with different polarity but same meaning) for Study 1. We tested the following six statements (all negation claim types):

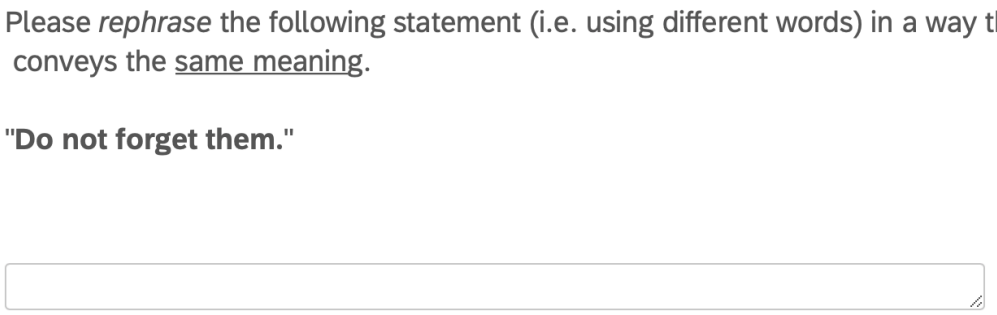
- We must not accept this injustice.
- We must not tolerate this injustice.
- Do not forget them.
- Do not overlook them.
- War is not acceptable.
- War is not OK.

We asked participants to rephrase them (i.e., rewrite the statement using different words) in a way that conveyed the same meaning, and how difficult it was to come up with such statement. Six hundred and three participants were recruited on Amazon Mechanical Turk, in exchange for monetary compensation. One hundred sixty-five participants performed the experimental task incorrectly (e.g., rephrased the statement conveying a different meaning from the one of the given statement) and were excluded from the analysis. Four hundred and thirty-eight participants (46.58% female, 0% other, $M_{Age} = 39.28$, $SD_{Age} = 11.51$) completed the task correctly. Participants saw only one of the six statements, randomly assigned.

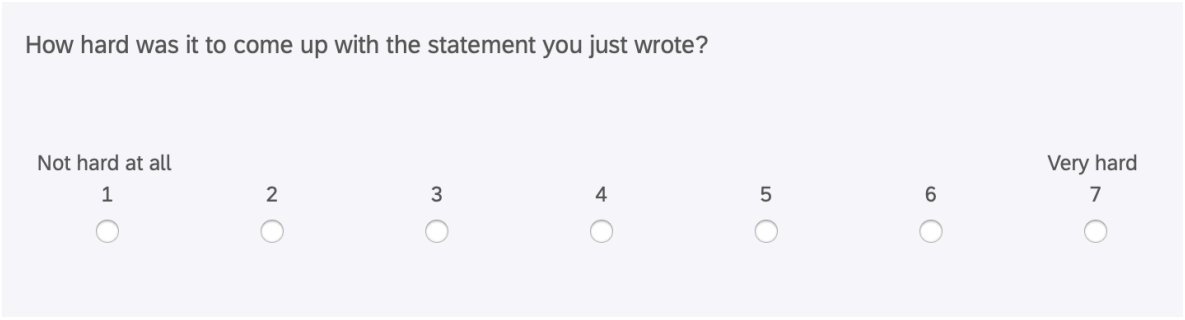
We categorized statements rephrased as negations (i.e., including a 'not') and with a synonym of the given word as "Uni-polar", and statements rephrased as affirmations (i.e., not including a 'not') and with an antonym of the given word as "Bi-polar". A chi-square test revealed that

participants rephrased ‘We must not accept this injustice’ in a bi-polar fashion as often as they did for ‘We must not accept this injustice’ ($\chi^2(1) = 2.10$, $p = .144$). A second chi-square test revealed that participants rephrased ‘War is not acceptable’ in a bi-polar fashion as often as they did for ‘War is not OK’ ($\chi^2(1) = 1.50$, $p = .216$). A third chi-square test, instead, revealed that participants rephrased ‘Do not forget them’ in a bi-polar fashion more often (74.2%) than they did for ‘Do not overlook them’ (24.1%; $\chi^2(1) = 43$, $p < .001$). For this reason, we selected the statements ‘Do not forget them’ and ‘Do not overlook them’ as the stimuli to use in Study 1.

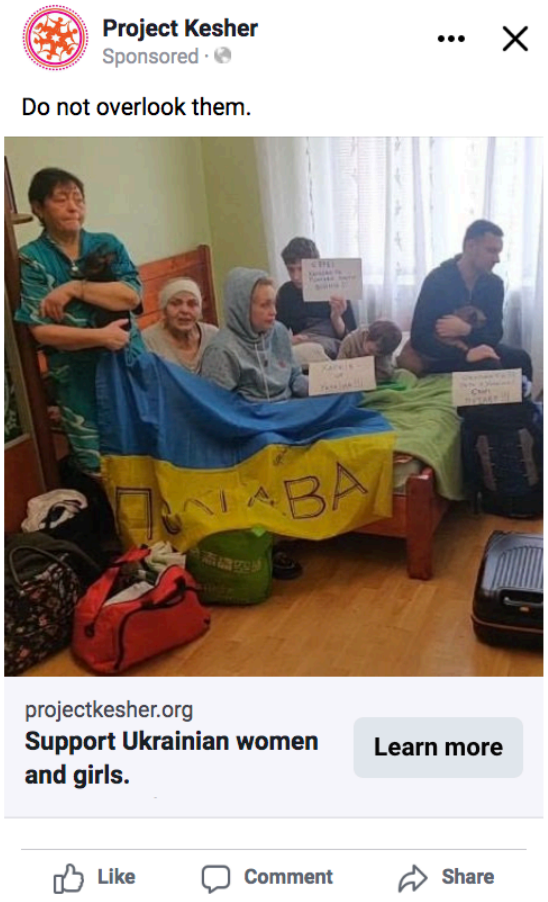
EXAMPLE OF REPHRASING QUESTION IN THE PRE-TEST OF STUDY 1

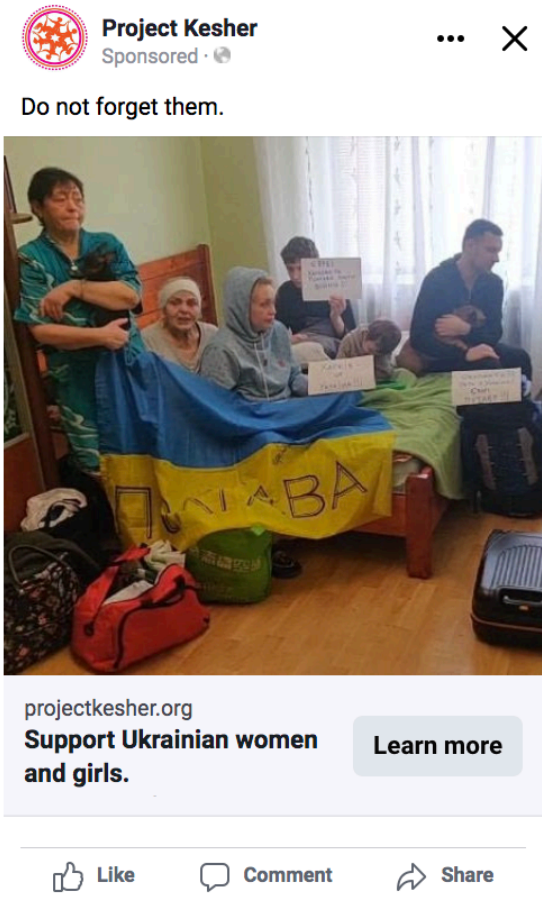


EXAMPLE OF DIFFICULTY QUESTION IN THE PRE-TEST OF STUDY 1



STIMULI FOR STUDY 1





APPENDIX

DESCRIPTION PRE-TEST 1 OF STUDY 2

Participants were recruited at the behavioral laboratory of a large west coast university. We pre-set the sample size as two days of data collection in the lab. One hundred and thirty undergraduates participated and were compensated with course credit. Four participants did not perform the experimental task correctly and were therefore excluded from the analysis leaving a sample of one hundred and twenty-six participants (54.76% female, $M_{Age} = 21.39$, $SD_{Age} = 2.99$).

All participants were shown twenty-nine single-word concepts (see section 2. of Web Appendix). For each one, they were asked to write down an antonym (if possible) and to rate the difficulty of finding an antonym on a scale from 1 (not at all) to 7 (extremely).

We z-scored each individual's difficulty ratings and used these normalized scores to compute the mean difficulty per concept across participants. Antonyms were defined as "correct" if they aligned with those listed in the Merriam-Webster online dictionary. As anticipated, across concepts there was a significant negative correlation ($r = -0.89$, $t(27) = -9.88$, $p < .001$) between the percentage of correctly identified antonyms and the mean perceived difficulty of retrieving those antonyms (Figure 1).

Based on these findings, we selected fifteen concepts, seven bi-polar and eight uni-polar (Figure 1; Table 1) to use in Study 1. We selected the bi-polar concepts based on the following criteria: more than 80% of respondents could find a correct antonym, and the mean z-scored difficulty of doing so was lower than -0.25. We selected the uni-polar concepts according to whether less than 20% of respondents could find a correct antonym, and the mean z-scored difficulty of doing so was greater than 0.25 (see Figure 1).

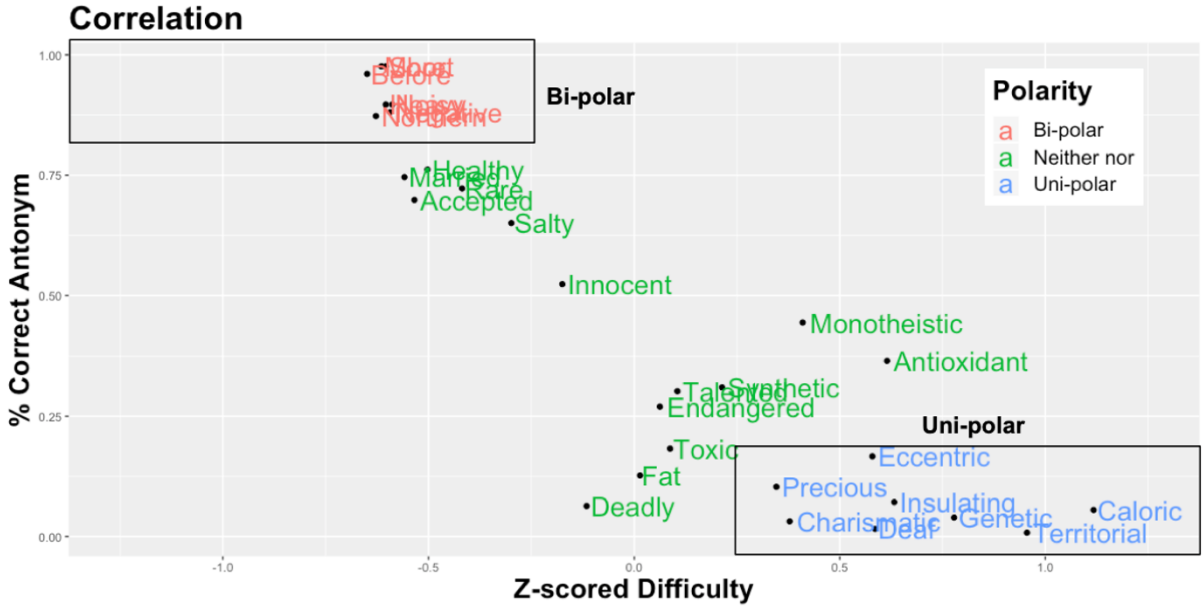


FIGURE 1. CORRELATION OF PERCENTAGE OF CORRECT ANTONYMS FOUND AND MEAN Z-SCORED DIFFICULTY PER CONCEPT IN PRE-TEST 1 OF STUDY 2

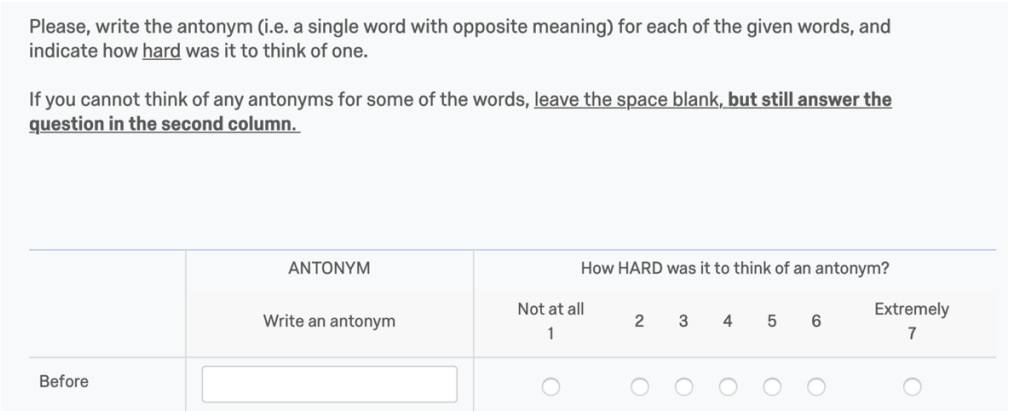
Bi-polar	Uni-polar
Before	Caloric
More	Deaf
Heavy	Precious
Negative	Territorial
Short	Charismatic
Northern	Insulating
Noisy	Eccentric
	Genetic

TABLE 1. CONCEPTS SELECTED FROM PRE-TEST 1 OF STUDY 2

CONCEPTS TESTED AND SELECTED IN PRE-TEST 1 OF STUDY 2

Concepts tested in the Pre-Test	% Correct Antonym	Z-scored Difficulty	Clear Polarity	CONCEPT
Before	96.0%	-0,64899602	Yes	Bi-polar
Accepted	69.8%	-0,53402035	No	
More	97.6%	-0,61484066	Yes	Bi-polar
Monotheistic	44.4%	0,41007327	No	
Innocent	52.4%	-0,17436296	No	
Heavy	94.1%	-0,60429888	Yes	Bi-polar
Negative	88.1%	-0,5937571	Yes	Bi-polar
Short	97.6%	-0,60429888	Yes	Bi-polar
Synthetic	31.0%	0,21444597	No	
Northern	87.3%	-0,6277438	Yes	Bi-polar
Rare	72.2%	-0,41806078	No	
Fat	12.7%	0,01415216	No	
Healthy	76.2%	-0,50239501	No	
Deadly	6.3%	-0,11586311	No	
Caloric	5.6%	1,1175250	Yes	Uni-polar
Deaf	1.6%	0,58692216	Yes	Uni-polar
Toxic	18.6%	0,08794462	No	
Precious	10.3%	0,34631659	Yes	Uni-polar
Territorial	0.8%	0,95588443	Yes	Uni-polar
Charismatic	3.2%	0,37819493	Yes	Uni-polar
Noisy	89.7%	-0,58878138	Yes	Bi-polar
Married	74.6%	-0,55861784	No	
Endangered	27.0%	0,06295357	No	
Salty	65.1%	-0,29833428	No	
Insulating	7.1%	0,6326032	Yes	Uni-polar
Eccentric	16.7%	0,58009109	Yes	Uni-polar
Genetic	4.0%	0,7784452	Yes	Uni-polar
Antioxidant	36.5%	0,61503357	No	
Talented	30.2%	0,10551425	No	

INSTRUCTIONS AND QUESTIONS PRE-TEST 1 OF STUDY 2


Please, write the antonym (i.e. a single word with opposite meaning) for each of the given words, and indicate how hard was it to think of one.

If you cannot think of any antonyms for some of the words, leave the space blank, **but still answer the question in the second column.**

	ANTONYM	How HARD was it to think of an antonym?					
	Write an antonym	Not at all 1	2	3	4	5	6	Extremely 7
Before		○	○	○	○	○	○	○

STATEMENTS TESTED IN PRE-TEST 2 OF STUDY 2

Field	Concept	Polarity	Claim Type	Statements	Truthfulness
Do you have extremely basic knowlwdge in American history?	Before	Bi-polar	Negation	The Bill of Rights was not ratified before the Constitution	T
Do you have extremely basic knowlwdge in American history?	Before	Bi-polar	Negation	Donald Trump was not elected before Barack Obama	T
Do you have extremely basic general knowledge?	More	Bi-polar	Negation	King Henry VIII did not have more than 4 wives	F
Do you have extremely basic general knowledge?	More	Bi-polar	Negation	Turtles can not live more than 100 years	F
Do you have extremely basic knowledge in chemistry?	Negative	Bi-polar	Negation	The charge of protons is not negative	T
Do you have extremely basic knowledge in chemistry?	Negative	Bi-polar	Negation	Magnetic field lines do not start from the negative pole	T
Do you have extremely basic knowledge in natural sciences?	Heavy	Bi-polar	Negation	An adult lion is not heavier than an adult cheeta	F
Do you have extremely basic knowledge in natural sciences?	Heavy	Bi-polar	Negation	1 dm^3 of lead is not heavier than 1 dm^3 of steel	F
Do you have extremely basic knowledge in cinema?	Short	Bi-polar	Negation	Schindler's List is not shorter than The Last Samurai	T
Do you have extremely basic knowledge in cinema?	Short	Bi-polar	Negation	Gone with the Wind is not shorter than The Notebook	T
Do you have extremely basic knowledge in geography?	Northern	Bi-polar	Negation	Italy is not in the Northern hemisphere	F
Do you have extremely basic knowledge in geography?	Northern	Bi-polar	Negation	Chicago is not in the Northern part of Illinois	F
Do you have extremely basic knowledge in acoustics?	Noisy	Bi-polar	Negation	Explosions in outer space are not noisy	T
Do you have extremely basic knowledge in acoustics?	Noisy	Bi-polar	Negation	Ultraviolet rays are not noisy	T
Do you have extremely basic knowledge about food properties?	Caloric	Uni-polar	Negation	Cheddar cheese is not a caloric food	F
Do you have extremely basic knowledge about food properties?	Caloric	Uni-polar	Negation	Avocado is not a caloric fruit	F
Do you have extremely basic knowledge about animal develpment?	Deaf	Uni-polar	Negation	Herbivores in the savannah are not deaf when they are born	T
Do you have extremely basic knowledge in music history?	Deaf	Uni-polar	Negation	Wolfgang Amadeus Mozart did not become deaf	T
Do you have extremely basic knowledge in jewelery?	Precious	Uni-polar	Negation	Platinum is not a precious metal	F
Do you have extremely basic knowledge in jewelery?	Precious	Uni-polar	Negation	Ruby is not a precious stone	F
Do you have extremely basic knowledge in zoology?	Territorial	Uni-polar	Negation	Sharks are not territorial animals	T
Do you have extremely basic knowledge in zoology?	Territorial	Uni-polar	Negation	Snails are not territorial animals	T
Do you have extremely basic knowledge in Roman history?	Charismatic	Uni-polar	Negation	Julius Caesar did not have a charismatic personality	F
Do you have extremely basic knowledge in Roman history?	Charismatic	Uni-polar	Negation	Augustus was not a charismatic emperor	F
Do you have extremely basic knowledge in material science?	Insulating	Uni-polar	Negation	Glass is not an insulating material	T
Do you have extremely basic knowledge in material science?	Insulating	Uni-polar	Negation	Stone is not an insulating material	T
Do you have extremely basic knowledge in art history?	Eccentric	Uni-polar	Negation	Salvador Dalí did not have an eccentric personality	F
Do you have extremely basic knowledge in art history?	Eccentric	Uni-polar	Negation	Pablo Picasso did not have an eccentric personality	F
Do you have extremely basic knowledge in genetics?	Genetic	Uni-polar	Negation	Hypochondria is not a genetic condition	T
Do you have extremely basic knowledge in genetics?	Genetic	Uni-polar	Negation	Traumas are not genetic conditions	T
Do you have extremely basic knowlwdge in American history?	Before	Bi-polar	Affirmation	The American Revolution happened before the First World War	T
Do you have extremely basic knowlwdge in American history?	Before	Bi-polar	Affirmation	John F. Kennedy was elected before Ronald Regan	T
Do you have extremely basic general knowledge?	More	Bi-polar	Affirmation	Permafrost accounts for more than 20% of total water on Earth	F
Do you have extremely basic general knowledge?	More	Bi-polar	Affirmation	William Shakespeare lived more than 90 years	F
Do you have extremely basic knowledge in chemistry?	Negative	Bi-polar	Affirmation	The charge of electrons is negative	T
Do you have extremely basic knowledge in chemistry?	Negative	Bi-polar	Affirmation	In a magnet the southern pole is the negative pole	T
Do you have extremely basic knowledge in natural sciences?	Heavy	Bi-polar	Affirmation	An adult giraffe is heavier than an adult elephant	F
Do you have extremely basic knowledge in natural sciences?	Heavy	Bi-polar	Affirmation	One liter of oil is heavier than one liter of water	F
Do you have extremely basic knowledge in cinema?	Short	Bi-polar	Affirmation	Gladiator is shorter than Titanic	T
Do you have extremely basic knowledge in cinema?	Short	Bi-polar	Affirmation	A Beautiful Mind is shorter than Dances with Wolves	T
Do you have extremely basic knowledge in geography?	Northern	Bi-polar	Affirmation	Argentina is in the Northern hemisphere	F
Do you have extremely basic knowledge in geography?	Northern	Bi-polar	Affirmation	Miami is in the Northern part of Florida	F
Do you have extremely basic knowledge in acoustics?	Noisy	Bi-polar	Affirmation	Thunders are noisy	T
Do you have extremely basic knowledge in acoustics?	Noisy	Bi-polar	Affirmation	Electromagnetic waves can be noisy	T
Do you have extremely basic knowledge about food properties?	Caloric	Uni-polar	Affirmation	Celery is a caloric food	F
Do you have extremely basic knowledge about food properties?	Caloric	Uni-polar	Affirmation	Coffee is a caloric drink	F
Do you have extremely basic knowledge about animal develpment?	Deaf	Uni-polar	Affirmation	Dogs are deaf when they are born	T
Do you have extremely basic knowledge in music history?	Deaf	Uni-polar	Affirmation	Ludwig van Beethoven bacame deaf	T
Do you have extremely basic knowledge in jewelery?	Precious	Uni-polar	Affirmation	Copper is a precious metal	F
Do you have extremely basic knowledge in jewelery?	Precious	Uni-polar	Affirmation	Onyx is a precious stone	F
Do you have extremely basic knowledge in zoology?	Territorial	Uni-polar	Affirmation	Lions are territorial animals	T
Do you have extremely basic knowledge in zoology?	Territorial	Uni-polar	Affirmation	Wolves are territorial animals	T
Do you have extremely basic knowledge in Roman history?	Charismatic	Uni-polar	Affirmation	Commodus was a charismatic emperor	F
Do you have extremely basic knowledge in Roman history?	Charismatic	Uni-polar	Affirmation	Nero was a charismatic emperor	F
Do you have extremely basic knowledge in material science?	Insulating	Uni-polar	Affirmation	Wood is an insulating material	T
Do you have extremely basic knowledge in material science?	Insulating	Uni-polar	Affirmation	Wool is an insulating material	T
Do you have extremely basic knowledge in art history?	Eccentric	Uni-polar	Affirmation	Michelangelo had an eccentric personality	F
Do you have extremely basic knowledge in art history?	Eccentric	Uni-polar	Affirmation	Vincent Van Gogh had an eccentric personality	F
Do you have extremely basic knowledge in genetics?	Genetic	Uni-polar	Affirmation	Anemia is a genetic disorder	T
Do you have extremely basic knowledge in genetics?	Genetic	Uni-polar	Affirmation	Down syndrome is a genetic disorder	T

DESCRIPTION OF PRE-TEST 2 OF STUDY 2

Our framework relates to situations in which consumers compare the statements they read—or hear—to their pre-existing beliefs. Holding a pre-existing belief about a statement leads to meaningful true-or-false judgments, thereby defining how that statement fits into the statement encoding matrix (Figure 1 in the paper). This will further allow us to classify the statement as either engaging SPT or Fusion. Thus it is important to select stimuli that are likely to relate to an existing belief (regardless of whether the individual's belief is accurate and/or correct).

Participants were recruited at the behavioral laboratory of a large west coast university. We pre-set the sample size as one week of data collection in the lab. Three hundred and nineteen undergraduates (40.44% female, $M_{Age} = 21.15$, $SD_{Age} = 2.43$) participated and were compensated with course credit. For each of the fifteen concepts selected in Pre-test 1, we created 4 Statement-type trials, reflecting two affirmation claim types and two negation claim types. This resulted in 60 Statement-type trials total. We also created 16 Question-type trials that were chosen to have answers that were likely to predict whether participants had pre-existing beliefs about the related Statements-type trials. For example, the answer (yes vs. no) to the Question-type trial "Do you have very basic knowledge in geography?" was meant to assess

the likely presence of a pre-existing belief about the truthfulness of the Statement-type trial "Italy is not in the Northern hemisphere". One Question-type trial served roughly four Statement-type trials. We aimed to select Question-type trials for which a 'yes' answer was predictive of whether participants hold pre-existing beliefs about any related Statements-type trials.

Participants first were asked to answer the 16 Question-type trials, and then were asked to judge the veracity of the Statement-type trials. One possible methodological concern is that answering 16 Question-type trials and judging 60 Statement-type trials is overly effortful and/or fatiguing. Thus, we divided the 60 Statement-type trials into four groups—each covering all the words identified in Pre-test 1. So, each participant was presented with the 16 Question-type trials and 15 of the 60 Statement-type trials (i.e., one of these four groups of Statement-type trials, randomly assigned). Specifically, participants answered 'yes' or 'no' to the 16 Question-type trials, judged 15 Statement-type trials as 'true' or 'false', and rated their confidence in those judgments on a scale from 50 (completely uncertain) to 100 (completely certain).

For each of the 60 Statement-type trials, we regressed the judgment confidence ratings on participants' answer to the related knowledge Question-type trial (answer variable: yes=1, no=0). Based on these regressions, we selected the Statement-type trials for which confidence was related to reported existing knowledge (i.e., had a significant positive coefficient on the "answer variable"). This yielded 22 Statement-type trials.

For each of these 22 Statement-type trials, we only considered the observations of those who answered 'yes' to the related Question-type trial. Within these observations, we counted the number of participants who gave the lowest judgment confidence rating of 50 ("completely uncertain"). Finally, we selected the 12 Statement-type trials that spanned all types of polarity and claim types (i.e., 3 bi-polar affirmations, 3 bi-polar negations, 3 uni-polar affirmations, and 3 uni-polar negations) with the fewest "completely uncertain" judgments. The resulting 12 Statement-type trials (and related 11 Question-type trials) are reported in Table 2 in the paper.

KNOWLEDGE QUESTION IN PRE-TEST 2 OF STUDY 2

Do you have very basic knowledge in the following subjects?

	Yes	No
American History	○	○
Chemistry	○	○
Natural Sciences	○	○
Cinema	○	○
Geography	○	○
Acoustics	○	○
Food Properties	○	○
Animals Development	○	○
Music History	○	○
Jewelery	○	○
Zoology	○	○
Roman History	○	○
Material Science	○	○
Art History	○	○
Genetics	○	○
General Knowledge	○	○

EXAMPLE OF TRUE/FALSE QUESTION IN PRE-TEST 2 OF STUDY 2

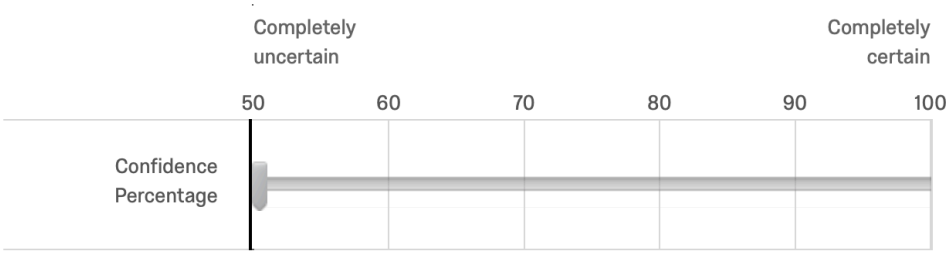
The Bill of Rights was not ratified before the Constitution

True ○ False ○

CONFIDENCE QUESTION IN PRE-TEST 2 OF STUDY 2

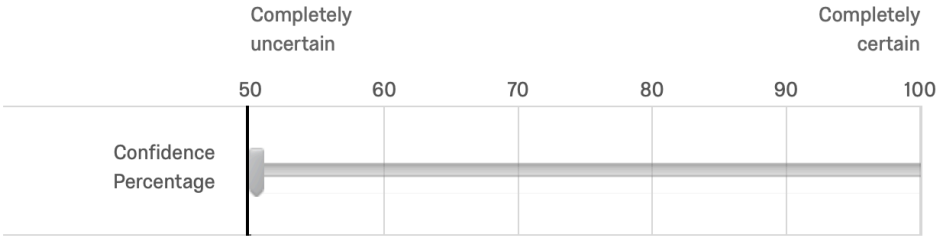
Display This Question:
If The Bill of Rights was not ratified before the Constitution True Is Selected

How confident are you that "The Bill of Rights was not ratified before the Constitution" is a true statement?

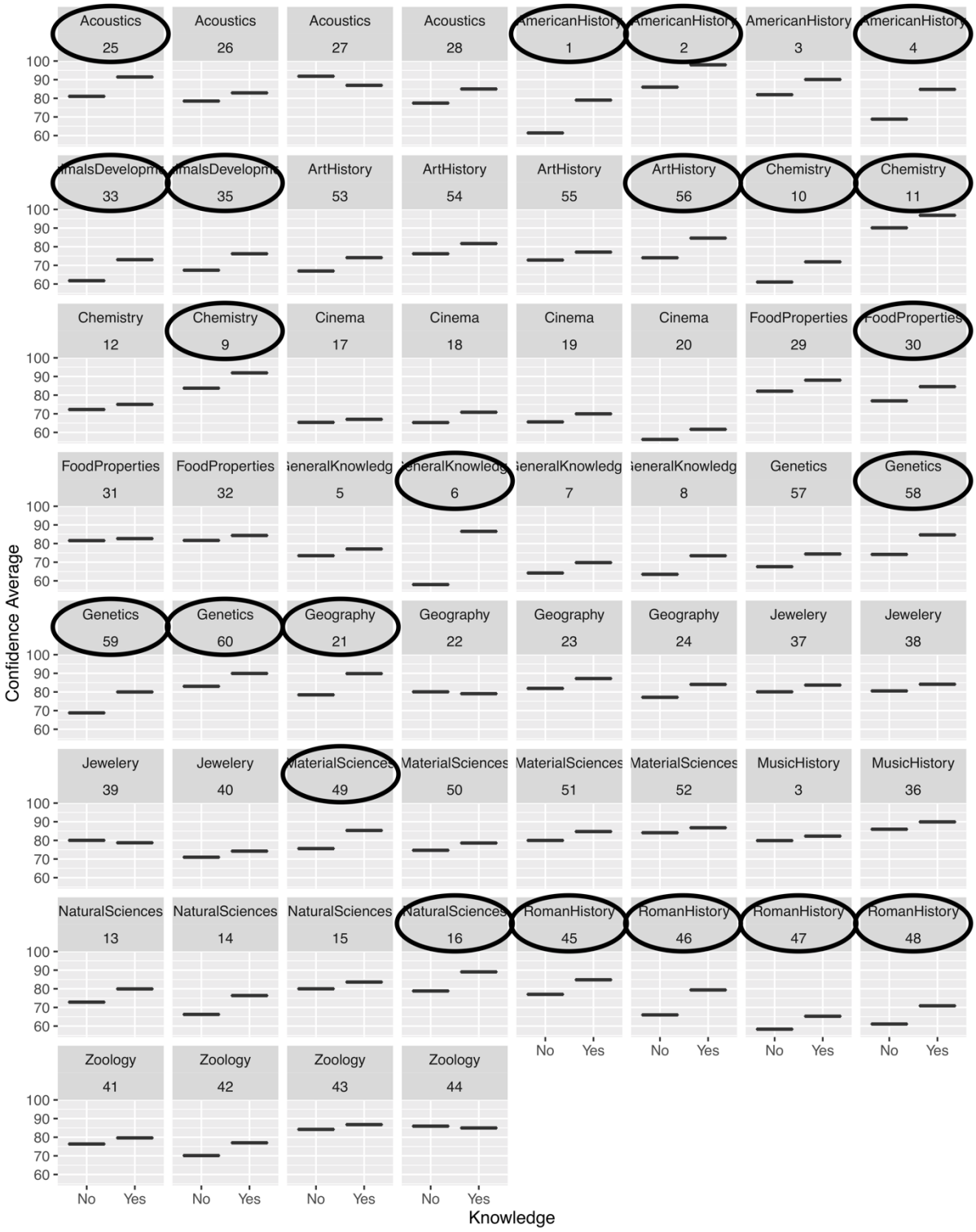


Display This Question:
If The Bill of Rights was not ratified before the Constitution False Is Selected

How confident are you that "The Bill of Rights was not ratified before the Constitution" is a false statement?

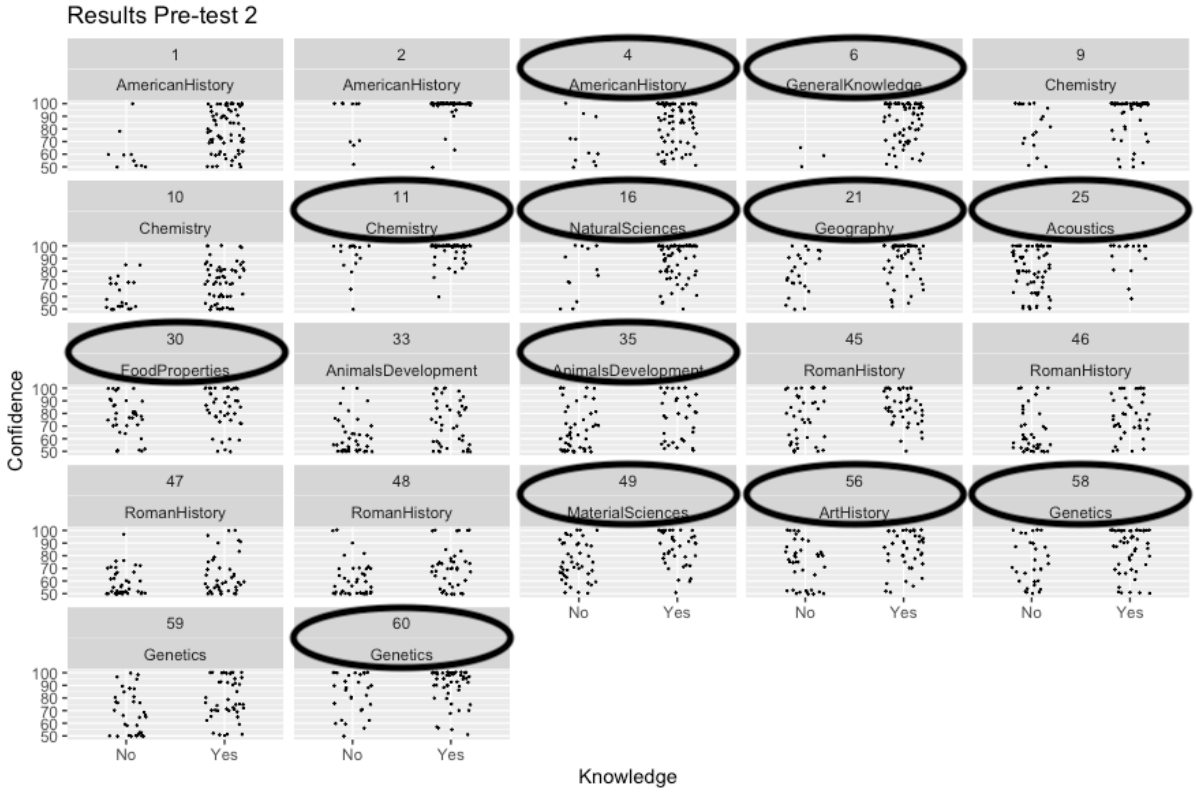


CONFIDENCE FOR EACH OF THE 60 STATEMENTS IN PRE-TEST 2 OF STUDY 2, GIVEN YES/NO ANSWER TO RELATIVE KNOWLEDGE QUESTION: 22 YIELDED SIGNIFICANT DIFFERENCE IN CONFIDENCE



CONFIDENCE DISTRIBUTION OF THE 22 SELECTED STATEMENTS IN PRE-TEST 2 OF STUDY 2, GIVEN A YES/NO ANSWER TO THE RELATIVE KNOWLEDGE QUESTION

We selected the 12 statements (3 uni-polar affirmations, 3 bi-polar affirmations, 3 uni-polar negations, and 3 bi-polar negations) with the fewest observation at "50 – completely uncertain" (circled).



STIMULI SELECTED FOR STUDY 2

Stimuli arising from Pre-test 2 of Study 2. Note we have two types of stimuli: the Statement-type trials (***S***; stimuli for the analysis) and the Question-type trials (***Q***; stimuli to define exclusions).

	Affirmations	**Negations**
Bi-polar	***S****:* John F. Kennedy was elected before Ronald Regan ***Q***: Do you have very basic knowledge in American history?	***S***: Turtles can not live more than 100 years ***Q***: Do you have very basic general knowledge?
	S*:* The charge of electrons is negative ***Q***: Do you have very basic knowledge in chemistry?	***S****:* Italy is not in the Northern hemisphere ***Q***: Do you have very basic knowledge in geography?
	S*:* One liter of oil is heavier than one liter of water ***Q***: Do you have very basic knowledge in natural sciences?	***S****:* Explosions in outer space are not noisy ***Q***: Do you have very basic knowledge in acoustics?
Uni-polar	***S****:* Dogs are deaf when they are born ***Q***: Do you have very basic knowledge about animal development?	***S****:* Avocado is not a caloric fruit ***Q***: Do you have very basic knowledge about food properties?
	S*:* Vincent Van Gogh had an eccentric personality ***Q***: Do you have very basic knowledge in art history?	***S****:* Glass is not an insulating material ***Q***: Do you have very basic knowledge in material science?
	S: Down Syndrome is a genetic disorder ***Q***: Do you have very basic knowledge in genetics?	***S****:* Traumas are not genetic conditions ***Q****:* Do you have very basic knowledge in genetics?

INSTRUCTIONS FOR STUDY 2

First instructions screen

```
You are going to answer 11 questions, each one on a separate screen.
Please, answer the 11 questions as you deem fit.

To answer the questions, press the key Y, if you want to answer YES,
or the key N, if you want to answer NO.

Please, press any key to go to the next page and read the first question.
```

Second instructions screen (after having answered the 11 questions)

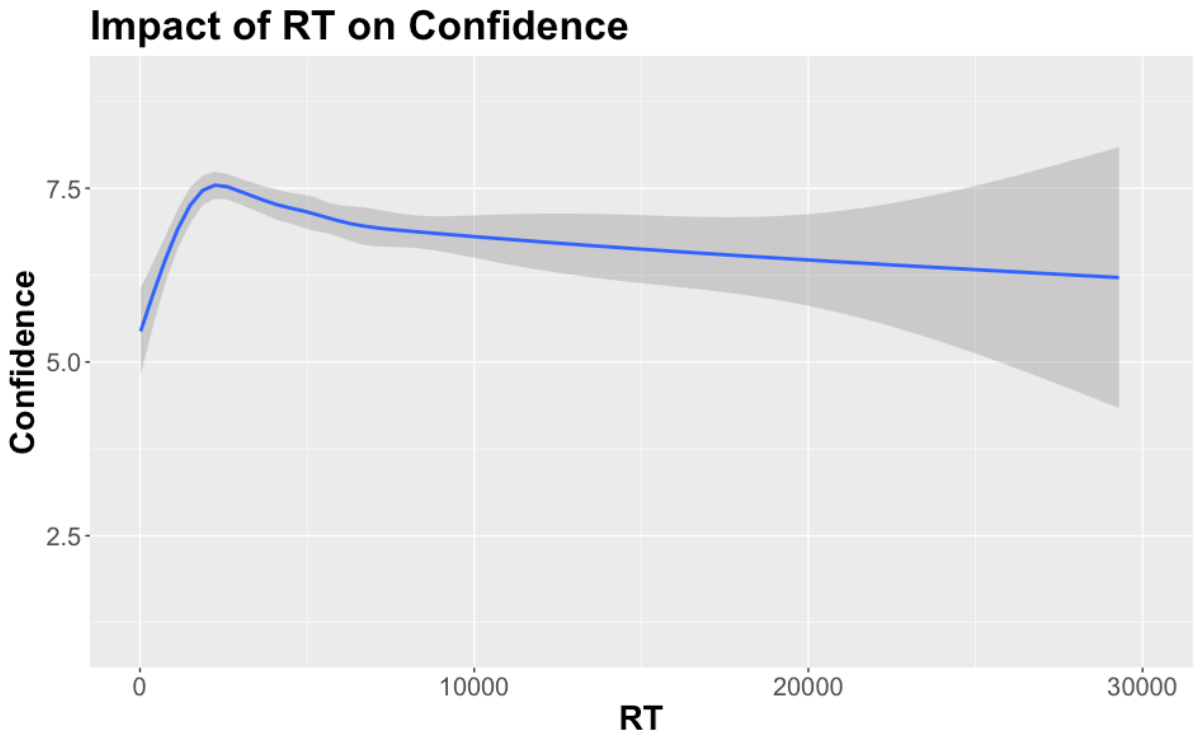
```
In the next page, you are going to read the instructions of this study.
We ask you to READ THE INSTRUCTIONS EXTREMELY CAREFULLY
and move on to the experiment only once you are  sure you are perfectly
familiar with the instructions, because you won't get the chance to
read them again during the study, and you can get a $10 Amazon.com gift
certificate based on your performance.

Please, press any key to go to the next page and read the instructions.
```

Third instructions screen (right after the second instructions screen)

```
INSTRUCTIONS


In this study, you are going to read 20 statements. Each statement will be displayed on a
different screen.
Every time you read a statement, you have to press either the key T on the keyboard, if you
think the statement is TRUE, or the key F, if you think the statement is FALSE.

After you judge each statement as TRUE or FALSE, a new screen will appear with the question
'How confident are you that you correctly judged the statement you just read as TRUE or
FALSE?'
In order to answer this question, you have to press any number on the keyboard from 1 to 9
(where 1=Completely uncertain about my answer, and 9=Completely certain about my answer).

Note that we have developed an algorithm that will identify the top 5 participants who
correctly judge the greatest number of statements as TRUE or FALSE, with the greatest level
of accuracy in assessing their confidence, and in the shortest amount of time. These 5
participants will be contacted by the end of the month and will receive a $10 Amazon.com gift
certificate as a prize.

When you feel familiar with the rules of the study, please press any key to start the
experiment.
```

THIRD PRE-REGISTERED MODEL IN STUDY 2

We ran a linear fixed effects model regressing judgment confidence on judgment RT, with random effects for subjects and Statement-type trials. We found that increases in RT (i.e., using one additional millisecond to judge a Statement-type trial) corresponds to lower confidence ratings (b = -0.000064, $t(2083) = -5.32$, $p < .001$).



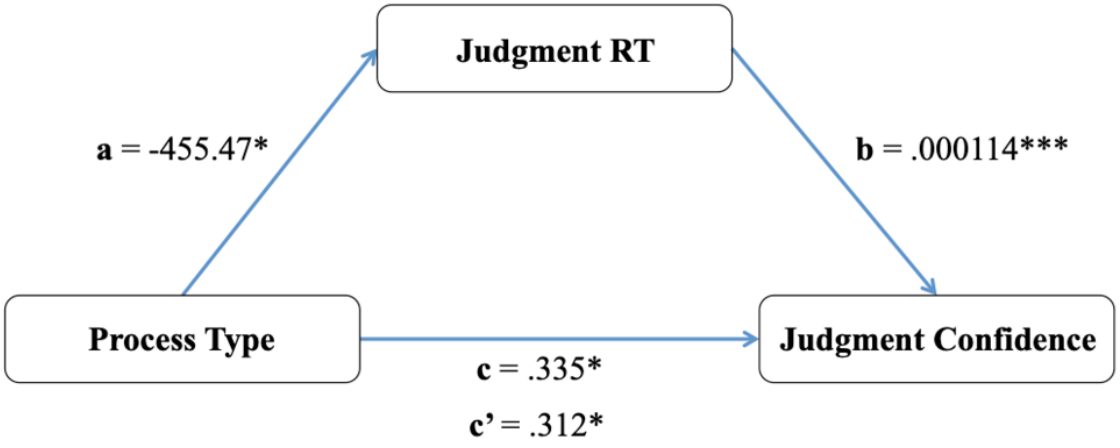
EXPLORATORY MODEL PRE-REGISTERED IN STUDY 2

We tested participant calibration, i.e., whether confidence relates meaningfully to performance. Across Statement-type trials, we considered judgment accuracy as whether participants correctly identified a Statement-type trial as true or false in the study. Using a mixed effects logistic model we regressed judgment accuracy on process type, with subjects and Statement-type trials as random effects. We find that Fusion has a less positive impact on accuracy compared to SPT (b = -1.51, z = -9.19, $p <$ .001).

MEDIATION ANALYSIS IN STUDY 2

Our findings demonstrate that judging a statement encoded via a Fusion process leads to both lower judgment confidence, and slower response times (RTs) than via SPT. While we measured RT as a manipulation check (i.e., an indication of whether participants engaged in the more effortful Fusion process), in research related to problem solving, RT is actually related to judgment confidence (Thompson et al., 2011). Furthermore, this relationship between RT and judgment confidence appears to exist independently of judgment accuracy (Ackerman and Zalmanov, 2012). Interestingly, past research has also found RT to be correlated with processing fluency. Processing fluency can be defined as 'the experience of processing ease', and it has been found to influence a variety of judgments, including perceptions of beauty (Reber, Schwarz and Winkielman, 2004), as well as judgments of truth (Reber and

Schwarz, 1999). These lines of research raise the question of whether our measure of RT is also capturing differences in fluency, and whether RT might mediate the effect of process type on judgment confidence. We conducted a mediation analysis to test this possibility, but it failed to show a relationship between process type, RT and confidence (see below). Thus, in this context, RT may indeed reflect other cognitive elements besides fluency, such as the additional semantic retrieval required for Fusion processes.



Indirect Effect = .027, **CI** = [-.021 , .069]

STIMULI FOR STUDIES 3A AND 3B

Suppose that your city has upcoming city-level General Elections. Imagine that you are attending a town hall meeting for voters to get to know a particular candidate, George Smith.

In many of these elections, it is common for companies to lobby candidates for preferential treatment or support for their industry. Recently, there have been rumors of stronger influences like direct payments in exchange for political favors. Despite the fact that Smith has worked in some of the industries implicated, he comes from a highly respected non-profit that operates transparently.

In response to a voter's question about outside influences, Smith talks about his motivations for running for office, and the goals he has for the city going forward. Smith concludes his answer by saying: "**During my term in office, I will not accept [tolerate] any forms of bribery.**"

JUDGMENT QUESTION IN STUDIES 3A AND 3B

Do you believe the candidate's statement (in bold) to be true or false?

True ○ False ○

ATTITUDE QUESTION FOLLOWING A TRUTH JUDGMENT IN STUDIES 3A AND 3B

Display This Question:
If Do you believe the candidate's statement (in bold) to be true or false? True Is Selected

To what extent do you agree with the following statement:

"This candidate is an ethical person"

Neither disagree nor agree		Somewhat agree		Moderately agree		Strongly agree
0	1	2	3	4	5	6
○	○	○	○	○	○	○

ATTITUDE QUESTION FOLLOWING A FALSE JUDGMENT IN STUDIES 3A AND 3B

Display This Question:
If Do you believe the candidate's statement (in bold) to be true or false? False Is Selected

To what extent do you agree with the following statement:

"This candidate is not an ethical person"

Neither disagree nor agree		Somewhat agree		Moderately agree		Strongly agree
0	1	2	3	4	5	6
○	○	○	○	○	○	○

MEANING QUESTION (BI-POLAR STATEMENT) IN STUDY 3B

What do you think "**During my term in office, I will not accept any form of bribery**" means?

MEANING QUESTION (UNI-POLAR STATEMENT) IN STUDY 3B

What do you think "**During my term in office, I will not tolerate any form of bribery**" means?

MANIPULATION CHECK BI-POLAR STATEMENT IN STUDIES 3A AND 3B

Write an antonym (i.e. a single word with opposite meaning) for the word '**accept**.'

If you cannot think of any antonyms, type 'none.'

MANIPULATION CHECK UNI-POLAR STATEMENT IN STUDY 3A AND 3B

Write an antonym (i.e. a single word with opposite meaning) for the word '**tolerate**.'

If you cannot think of any antonyms, type 'none.'

DESCRIPTION PRE-TEST 1 OF STUDY 4

To identify pairs of words with the same meaning and opposite polarity, the experimenters selected 3 synonyms for 8 "meanings", some we predicted would be bi-polar and some others uni-polar (for a total of 24 concepts). The aim was to determine the concept polarity of the 24 words by collecting data on people's ability to retrieve their antonyms and on the difficulty of doing so. In order to avoid antonym retrieval spillovers from a bi-polar word to its uni-polar synonym in a within-subjects design, we divided the 24 concepts into three separate groups with only one concept per meaning in each group.
Three hundred participants were recruited on Amazon Mechanical Turk and engaged in the study for monetary compensation. Thirty-two participants performed the experimental task incorrectly (e.g., reporting synonyms instead of antonyms) and were excluded from the analysis, leaving a sample of 268 participants (43.66% female, 0.37% other, $M_{Age} = 37.72$, $SD_{Age} = 11.53$). Participants were randomly assigned to judge one of the three pre-defined groups of concepts. For each of the eight words (concepts) that participants saw, they were asked to write down an antonym (if possible), to rate the difficulty of finding said antonym on a scale from 1 (not at all) to 7 (extremely), and to indicate whether they knew the meaning of each word (Yes, No, Not sure).
Out of the 8 "meanings", we identified 3 for which we had at least a bi-polar and a uni-polar word:

- Uninteresting (bi-polar) and Monotonous (uni-polar)
- Unsatisfactory (bi-polar) and Lame (uni-polar)
- Stale and Unoriginal (bi-polar) and Banal (uni-polar)

We excluded Banal (and its bi-polar counterparts) because less than half of the sample reported knowing its meaning. We were left with:

- Uninteresting (bi-polar) and Monotonous (uni-polar)
- Unsatisfactory (bi-polar) and Lame (uni-polar).

CONCEPTS TESTED AND SELECTED IN PRE-TEST 1 OF STUDY 4

Predicted Polarity	Concept	% CorrAntonym	Z.Difficulty	% KnowMeaning
Bi-polar	**Weak**	**98.9%**	**-0.40**	
Uni-polar	Shaky	71.4%	**-0.31**	
Uni-polar	Flimsy	52.4%	**-0.32**	
Bi-polar	**Uninteresting**	**74.2%**	**-0.29**	**100%**
Uni-polar	**Monotonous**	**5.5%**	**+0.36**	**87.9%**
Bi-polar	Boring	20.2%	**-0.38**	
Uni-polar	Lousy	57.0%	+0.05	
Bi-polar	Unsatisfactory	67.0%	**-0.31**	98.9%
Uni-polar	**Lame**	**3.6%**	**+0.18**	**98.8%**
Bi-polar	**Insignificant**	**84.9%**	**-0.31**	
Uni-polar	Trivial	51.6%	**+0.14**	
Bi-polar	Inconsequential	77.4%	+0.02	
Uni-polar	**Compelling**	**11.8%**	**+0.44**	
Bi-polar	Convincing	39.6%	+0.22	
Bi-polar	Persuasive	52.4%	+0.67	
Bi-polar	**Stale**	**91.4%**	-0.18	**100%**
Uni-polar	**Banal**	**12.1%**	**+1.26**	**48.4%**
Bi-polar	Unoriginal	70.2%	**-0.41**	100%
Uni-polar	Yucky	68.8%	-0.12	
Bi-polar	Unsavory	75.8%	-0.09	
Bi-polar	Unpalatable	**85.7%**	+0.30	
Bi-polar	**Difficult**	**98.9%**	**-0.37**	
Bi-polar	**Complex**	**94.5%**	**-0.39**	
Uni-polar	Arduous	71.4%	**+0.40**	

NOTE—Concepts in the same color were presented to the same group of participants.

Thresholds:

- Bi-polar: %CorrAntonym > 70%, Z.Difficulty < - 0.23
- Uni-polar: %CorrAntonym < 30%, Z.Difficulty > + 0.11

DESCRIPTION OF PRE-TEST 2 OF STUDY 4

The second pre-test was used to select a specific pair of stimuli (i.e., statements) for Study 4 that offered the strongest polarity of our central concepts. We tested four negation claim types containing the four concepts identified in Pre-test 1:

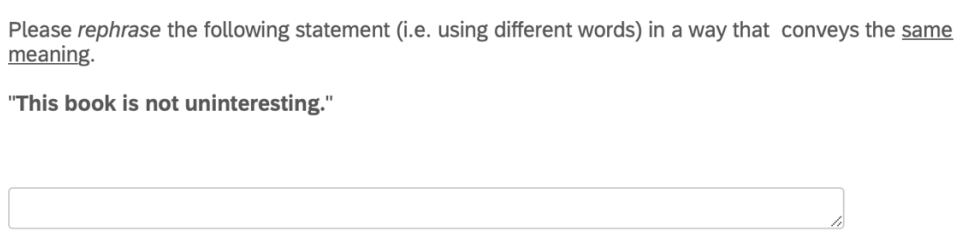
- This book is not uninteresting.
- This book is not monotonous.
- The experience was not unsatisfactory.
- The experience was not lame.

We asked participants to rephrase them (i.e., rewrite the statement using different words) in a way that conveyed the same meaning, and how difficult it was to come up with such statement.
Three hundred and twenty-one participants were recruited on Amazon Mechanical Turk, in exchange for monetary compensation. Sixty-six participants performed the experimental task incorrectly (e.g., rephrased the statement conveying a different meaning from the one of the given statement) and were

excluded from the analysis. Two hundred and fifty-five participants (47.45% female, 0.78% other, M_{Age} = 38.23, SD_{Age} = 11.81) completed the task correctly. Participants saw only one of the four statements, randomly assigned.

We categorized statements rephrased as negations (i.e., including a 'not') with a synonym of the given word as Uni-polar, and statements rephrased as affirmations (i.e., not including a 'not') with an antonym of the given word as Bi-polar. A chi-square test revealed that participants rephrased the statement with 'uninteresting' in a bi-polar fashion more often than they did for the statement with 'monotonous' ($\chi^2(1) = 24.90$, $p < .001$). A second chi-square test, instead, revealed that participants rephrased the statements with 'unsatisfactory' and 'lame' in a bi-polar fashion equally often ($\chi^2(1) = 1.25$, $p = .263$). For this reason, we selected the statements 'This book is not uninteresting' and 'This book is not monotonous' as the stimuli to use in Study 4.

EXAMPLE OF REPHRASING QUESTION IN PRE-TEST 2 OF STUDY 4



EXAMPLE OF DIFFICULTY QUESTION IN PRE-TEST 2 OF STUDY 4

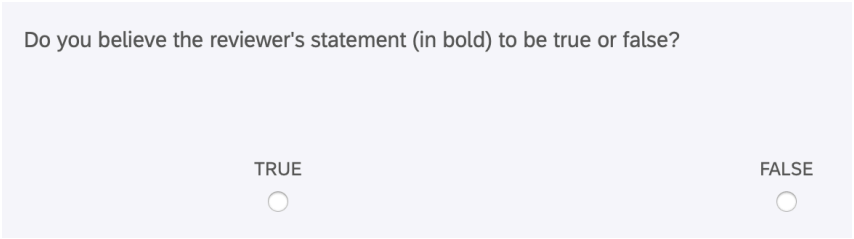
STIMULI FOR STUDY 4

Imagine you want to get yourself a good book to read in the coming weeks.

To buy one, you start scrolling through the options at an online bookstore. All of a sudden you see a cover that draws your attention. It's a novel with a catchy title, and it's exactly the genre you like. You take a closer look at the book description, and you learn it is the first novel from an up-and-coming writer.

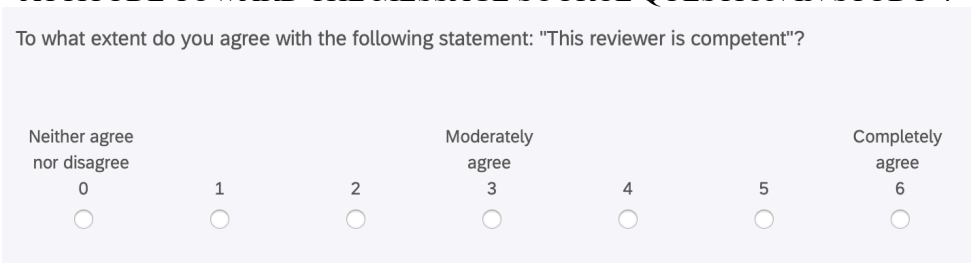
Under the cover picture, there is a review. The reviewer seems to like the book. They gave it a positive rating (4 out of 5 stars) and discussed their impression on the plot and the writing style. During the review, they note: "**This book is not uninteresting [monotonous]**".

JUDGMENT QUESTION IN STUDY 4



ATTITUDE TOWARD THE MESSAGE SUBJECT QUESTION IN STUDY 4

ATTITUDE TOWARD THE MESSAGE SOURCE QUESTION IN STUDY 4



PARAMETERS TABLE FROM THE 2x2 ANOVA IN STUDY 4

Parameter	Sum_Squares	df	Mean_Square	F	p
(Intercept)	16657.69	1	16657.69	10599.19	< .001
ProcessType	22.23	1	22.23	14.14	< .001
AttitudeType	23.09	1	23.09	14.69	< .001
ProcessType:AttitudeType	11.70	1	11.70	7.44	0.006
Residuals	1640.75	1044	1.57		

PARTIAL ETA SQUARED TABLE FROM THE 2x2 ANOVA IN STUDY 4

Parameter	Eta2 (partial)	95% CI
ProcessType	0.01	[0.00, 1.00]
AttitudeType	0.01	[0.00, 1.00]
ProcessType:AttitudeType	7.08e-03	[0.00, 1.00]

PARTIAL COHEN'S F TABLE FROM THE 2x2 ANOVA IN STUDY 4

Parameter	Cohen's f (partial)	95% CI
ProcessType	0.12	[0.07, Inf]
AttitudeType	0.12	[0.07, Inf]
ProcessType:AttitudeType	0.08	[0.03, Inf]

REFERENCES

Abelson, Robert P. (1995), "Attitude Extremity," In R. E. Petty and J. A. Krosnick (Eds.), *Ohio State University series on Attitudes and Persuasion: Vol. 4. Attitude Strength: Antecedents and Consequences*, 25-41. Lawrence Erlbaum Associates.

Ackerman, Rakefet, and Hagar Zalmanov (2012), "The Persistence of Fluency-Confidence Association in Problem Solving," *Psychonomic Bulletin and Review*, (19), 1187-1192.

Bassili, John N. (1996), "Meta-judgmental versus Operative Indexes of Psychological Attributes: The Case of Attitude Strength," *Journal of Personality ads Social Psychology*, 71(4), 637-653.

Fazio, Russell H., and Mark P. Zanna (1978), "Attitudinal Qualities Relating to the Strength of the Attitude-Behavior Relationship," *Journal of Experimental Social Psychology*, 14(4), 398-408.

Festinger, Leon (1950), "Informal Social Communication," *Psychological Review*, 57, 271-282.

Festinger, Leon (1954), "A Theory of Social Comparison Processes," *Human Relations*, 7, 117-140.

Fishbein, Martin, and Icek Ajzen (1975), *Belief, Attitude, Intention, and Behavior*, Reading, MA: Addison-Wesley.

Gilbert, Daniel T. (1991), "How Mental Systems Believe," *American Psychologist*, 46(2), 107-119.

Gilbert, Daniel T., Douglas S. Krull, and Patrick S. Malone (1990), "Unbelieving the Unbelievable: Some Problems in the Rejection of False Information," *Attitudes and Social Cognition*, 59(4), 601-613.

Grant, Susan J., Prashant Malaviya, and Brian Sternthal (2004), "The Influence of Negation on Product Evaluations," *Journal of Consumer Research*, 31, 583-591.

Gross, Sharon R., Rolf Holtz, and Norman Miller (1995), "Attitude Certainty," *Attitude strength: Antecedents and consequences*, 4, 215-245.

Karmarkar, Uma R., and Zakary L. Tormala (2010), "Believe Me, I Have No Idea What I'm Talking About: The Effects of Source Certainty on Consumer Involvement and Persuasion," *Journal of Consumer Research*, 36, 1033-1049.

Mayo, Ruth, Yaacov Schul, and Eugene Burnstein (2004), " "I am not guilty" vs. "I am innocent": Successful Negation May Depend on the Schema Used for its Encoding," *Journal of Experimental Social Psychology*, 40, 433-449.

McFerran, Brent, Sarah G. Moore, and Grant Packard (2019), "How Should Companies Talk to Customers Online?," *MIT Sloan Management Review*, 60(2), 68-71.

Orazi, Davide C., and Allen C. Johnston (2020), "Running Field Experiments Using Facebook Split Test," *Journal of Business Research*, 118, 189-98.

Packard, Grant, and Jonah Berger (2017), "How Language Shapes Word of Mouth's Impact," *Journal of Marketing Research*, 54, 572-588.

Packard, Grant, and Jonah Berger (2020a), "Thinking of You: How Second-Person Pronouns Shape Cultural Success," *Psychological Science*, 31(4), 397-407.

Packard, Grant, and Jonah Berger (2020b), "How Concrete Language Shapes Customer Satisfaction," *Journal of Consumer Research*, forthcoming.

Packard, Grant, Sarah G. Moore, and Brent McFerran (2018), "(I'm) Happy to Help (You): The Impact of Personal Pronoun Use in Customer-Firm Interactions," *Journal of Marketing Research*, 55, 541-555.

Petty, Richard E., and John T. Cacioppo (1986), "The Elaboration Likelihood Model of Persuasion," In *Communication and Persuasion* (pp. 1-24). Springer, New York, NY.

Petty, Richard E., Pablo Briñol, and Zakary L. Tormala (2002), "Thought Confidence as a Determinant of Persuasion: The Self-Validation Hypothesis," *Journal of Personality and Social Psychology*, 82(5), 722-741.

Petty, Richard E., and Duane T. Wegener (1998), "Attitude Change: Multiple Roles for Persuasion Variables," In *The Handbook of Social Psychology*, 4th ed. Daniel Gilbert, Susan Fiske, and Gardner Lindzey, New York: McGraw-Hill, 323-390.

Priester, Joseph R., and Richard E. Petty (1995), "Source Attributions and Persuasion: Perceived Honesty as a Determinant of Message Scrutiny," *Personality and Social Psychology Bulletin*, 21(6), 637-654.

Reber, Rolf, and Norbert Schwarz (1999), "Effects of Perceptual Fluency on Judgments of Truth," *Consciousness and Cognition*, (8), 338-342.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman (2004), "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?," *Personality and Social Psychology Review*, 8(4), 364-382.

Rosenblum, Michael, Juliana Schroeder, and Francesca Gino (2019), "Tell It Like It Is: When Politically Incorrect Language Promotes Authenticity," *Journal of Personality and Social Psychology*, 119(1), 1-29.

Thompson, Valerie A., Jamie A. Prowse Turner, Gordon Pennycook (2011), "Intuition, Reason, and Metacognition," *Cognitive Psychology*, 63(3), 107-140.

Tormala, Zakary L., and Richard E. Petty (2002), "What Doesn't Kill Me Makes Me Stronger: The Effects of Resisting Persuasion on Attitude Certainty," *Journal of Personality and Social Psychology*, 83, 1298-1313.

Tormala, Zakary L., and Richard E. Petty (2004a), "Source Credibility and Attitude Certainty: A Metacognitive Analysis of Resistance to Persuasion," *Journal of Consumer Psychology*, 14(4), 427-442.

Tormala, Zakary L., and Richard E. Petty (2004b), "Resistance to Persuasion and Attitude Certainty: The moderating Role of Elaboration," *Personality and Social Psychology Bulletin*, 30(11), 1446-1457.

Tormala, Zakary L., and Derek D. Rucker (2007), "Attitude Certainty: A Review of Past Findings and Emerging Perspectives," *Social and Personality Psychology Compass*, 1(1), 469-492.

Tormala, Zakary L., and Derek D. Rucker (in press), "Attitude Change and Persuasion: Classic, Metacognitive, and Advocacy Perspectives," In L. R. Kahle, T. M. Lowery, and J. Huber (Eds.), *APA Handbook of Consumer Psychology*. American Psychological Association.