Political Behavior
https://doi.org/10.1007/s11109-024-09993-z

ORIGINAL PAPER

Couples Therapy for a Divided America: Assessing the Effects of Reciprocal Group Reflection on Partisan Polarization

Hannah Baron[1] · Robert A. Blair[2] · Donghyun Danny Choi[2] · Laura Gamboa[3] · Jessica Gottlieb[4] · Amanda Lea Robinson[5] · Steven C. Rosenzweig[6] · Megan M. Turnbull[7] · Emily A. West[8]

Accepted: 16 October 2024

Abstract

Overcoming America's deep partisan polarization poses a unique challenge: Americans must be able to sharply disagree on who should govern while agreeing on more fundamental democratic principles. We study one model of depolarization—*reciprocal group reflection*—inspired by marital counseling and implemented by a non-partisan non-governmental organization, Braver Angels. We randomly assigned undergraduates at four universities either to participate in a Braver Angels workshop or simply to complete three rounds of surveys. The workshops substantially reduced polarization according to explicit and implicit measures. They also increased participants' willingness to donate to programs aimed at depolarizing political conversations. These effects are consistent across partisan groups, though some dissipate over time. Using qualitative data, and building on contact and deliberative theories, we argue that depolarization is especially effective when it includes both informational and emotional components, such that citizens who are moved to empathize with outgroup members become more likely to internalize new information about them.

We thank Lindsey Boss and the 2020 Democratic Erosion summer fellows for excellent research assistance. We also thank Braver Angels leadership and moderators for making this study possible. The study received funding from the Scowcroft Institute of International Affairs at the Bush School of Government and Public Service at Texas A&M University; the Institute for Democratic Engagement and Accountability (IDEA) at Ohio State University; the Swearer Center and Department of Political Science at Brown University; and the Department of Political Science at University of Pittsburgh. We thank participants at Temple University's Political Science Behavioral Workshop and an EGAP Discussion Session for their feedback. We also thank Liliana Mason for the code to measure issue extremity. This study is pre-registered with EGAP as Design 20200217AB, available at https://osf.io/w8kpe. It received IRB approval from Boston University under Protocol 5373E, with reliance agreements at the additional universities. Data and replication files are available at https://doi.org/10.7910/DVN/ZDFX1V.

Extended author information available on the last page of the article

Published online: 28 January 2025

Keywords Affective polarization · American politics · Intergroup contact · Field experiments

Like a couple who remain responsible for their children no matter what happens to their own relationship, Reds and Blues cannot simply walk away from each other. Neither side can "divorce" and move to a different country.

–Bill Doherty, Braver Angels Co-Founder

Introduction

The American public is deeply polarized along partisan lines (Iyengar et al., 2019; Mason, 2018). Americans increasingly view those on the other side of the partisan divide as untrustworthy, unpatriotic, and uninformed (Haidt & Hetherington, 2012; Iyengar et al., 2012). While scholars debate the causal relationship between affective polarization and democratic norms (Broockman, Kalla and Westwood, 2020), research suggests it can undermine democracy in several ways. For one, it can weaken electoral accountability for undemocratic actions by driving citizens to prioritize partisan preferences over democratic principles (Graham & Svolik, 2020). It can harm democratic legitimacy and the quality of democratic governance by encouraging legislative gridlock and disincentivizing compromise (Hetherington & Rudolph, 2015; McCoy et al., 2018). And it can threaten important democratic attitudes and norms, such as tolerance for the opposition (Kingzette et al., 2021; Levitsky & Ziblatt, 2018).

We present results from an experimental evaluation of an intervention that aims to address one of democracy's unique challenges: preserving the legitimacy of opposing political preferences and encouraging tolerance among citizens who sharply disagree over who should govern. Despite a modest sample size, our results suggest that the intervention—which brought together groups of Democrats and Republicans for workshops based on practices in couples counseling that seek to elicit mutual vulnerability and compassion while working to improve the health of a marriage (Doherty, 2021)—reduced affective polarization after one month; the effect on a behavioral measure persists after six months[1]. In the workshops, participants mainly interact with members of their *own* group, reflecting on the attitudes, beliefs, and characteristics of co-partisans. Crucially, when the two groups enter into dialogue with one another, they do not attempt to persuade or reach compromise on particular issues. Instead, direct intergroup dialogue focuses on generating understanding and tolerance of "the other" and the motivations and experiences behind their beliefs.

This intervention represents a new approach to overcoming affective polarization that is related to—but distinct from—existing intergroup relations theories such as con-

[1] Our randomized design mitigates concerns that the reported effects are artifacts of selection bias, but one might. be concerned about a Type I error due to greater sampling variability with our relatively small sample. We discuss evidence mitigating against this concern in Supplement Appendix E alongside a discussion of power given effect sizes in the relevant literature.

tact and deliberative democracy. In the former, one of the primary goals is to bring groups together to encourage cooperation toward a common goal (Allport, 1954); in the latter, it is to help ideologically divided individuals reach compromise or consensus on particular issues (Cohen, 1989). While the intervention we study shares some features with both of these models, it does *not* aim for cooperation, consensus, or compromise. Instead, it promotes depolarization by encouraging listening and understanding, skills which may be important for the quality of deliberation (Dobson, 2012).

Partnering with Braver Angels, a non-governmental organization, we evaluated the effects of this model—which we term *reciprocal group reflection*—with a field experiment involving 169 undergraduate students on four college campuses. Students are a particularly relevant population for this study, and for programs aimed at reducing polarization more generally: not only do they represent the future of the Republican and Democratic parties, but analysts have expressed particular concern about increasing levels of polarization among college students who are coming of age at a time of partisan echo chambers in the home, on campus, and on social media (Glatter, 2017; Iyengar et al., 2019).

Our results suggest that the workshops reduced pre-specified implicit and explicit measures of affective polarization, though the effect on the explicit measure just misses the conventional 5% threshold for statistical significance. The workshops also increased behavioral support for depolarization, operationalized as participants' willingness to donate money to programs aimed at depolarizing political conversations among American youths. The effects on our explicit and implicit measures of polarization dissipate somewhat over time, but the effects on the behavioral measure persist after more than half a year.

To better understand the mechanisms underlying these effects, we recorded and coded transcripts of the workshop proceedings. Inductive analysis of these transcripts suggests that the intervention was successful because it combined emotional and informational components, building empathy with outgroup members such that participants became more willing to update their priors about the outgroup in response to new information. Two unique features of the model facilitated this process. First, the group-based nature of the perspective-getting exercises helped participants disassociate outgroup members from stereotypes, then re-associate new learnings about specific out-group participants with the outgroup more generally. Second, the reciprocal nature of the exchange generated mutual vulnerability between ingroup and outgroup members. We argue this emotional engagement facilitated the acceptance of new information.

Our study makes several contributions to the literature on polarization and democracy, as well as to efforts to reduce affective polarization and bolster democratic resilience. First, we describe and test a new model of depolarization—reciprocal group reflection—that combines emotional and informational mechanisms to reduce animosity between groups. This approach encourages participants to reevaluate both the out-group *and* the in-group, distinguishing it from existing approaches geared mainly toward inducing contact, promoting deliberative skills, highlighting commonalities, or correcting misinformation (Hartman et al., 2022). Furthermore, we systematically study the potential mechanisms at play with both quantitative and qualitative data, taking particular advantage of the unusually rich qualitative data we collected from transcripts and direct observations of the workshops.

Second, we assess the effect of the intervention on a broader range of outcomes than are usually available in such studies, including both explicit and implicit attitudes towards out-partisans and a measure of actual behavior. Existing research largely focuses on explicit attitudes alone (Hartman et al., 2022). Furthermore, our collection of midline and endline data over several months allows us to estimate longer-run effects than most studies, and to explicitly test for decay over time, providing crucial information about the feasibility of durably reducing partisan animosity.

Finally, this study represents a successful academic-practitioner partnership, an unusual arrangement in a literature where most studies evaluate interventions that were designed and implemented by researchers (Hartman et al., 2022). While findings from these studies are informative, it is unclear whether the interventions they evaluate are scalable or sustainable beyond the lifespan of the evaluation itself. Our partner Braver Angels, on the other hand, has facilitated thousands of workshops in recent years through a decentralized ("cellular") organizational structure, a model that other organizations such as schools, religious institutions, and community groups could easily adopt. We therefore provide evidence in support of a scalable and sustainable approach to depolarization.

Affective Polarization: How Might Americans "Depolarize"?

Affective polarization captures the extent to which citizens express antipathy towards members of other parties and/or affinity for members of their own. This is distinct from (though often correlated with) ideological polarization, whereby citizens hold increasingly extreme issue positions (Iyengar et al., 2019; Webster & Abramowitz, 2017). Beyond mere disagreement over policy, affective polarization manifests in dislike and distrust of members of the opposing party and affection towards members of one's own. Most of the literature on partisan polarization in the US understands affective polarization as the result of the internalization of partisanship as a social identity, which generates strong ingroup preferences and outgroup bias (Huddy et al., 2015; Iyengar et al., 2019; West & Iyengar, 2022).

Deliberation and Intergroup Contact to Reduce Polarization

A testament to the growing importance of the problem, there are now hundreds of US-based researchers and organizations that have developed, implemented, and sometimes studied a range of different depolarization interventions with diverse designs (Hartman et al., 2022). Interventions that target polarized or prejudicial attitudes can be categorized into two main approaches: strategies that address informational deficits and/or misperceptions, and strategies that address affective and emotional barriers to positive feelings toward the outgroup. Intergroup contact interventions to reduce prejudice across a range of social and political cleavages are intended to activate both informational and emotional pathways by addressing mutual ignorance and reducing fear in order to diminish negative attitudes toward an outgroup.

Working through informational pathways, some successful depolarization interventions rely on individuals updating beliefs about ingroup-outgroup differences or similarities (Wojcieszak & Warner, 2020) and correcting misperceptions about outgroup members (Ahler & Sood, 2018) to shift attitudes. Standard deliberative polling interventions share a similar theory of change but do not typically target affective depolarization, focusing more on changes in public policy attitudes than affect toward groups. Instead, individuals acquire new knowledge—by updating information about policies and about the beliefs and opinions that other citizens hold—which enables them to revise their own political opinions. Within moderated, small group discussions using balanced materials on policy issues, the deliberative process involves slowed-down thinking and careful consideration of public policy trade-offs[2]. Importantly, dialogue is supposed to take place among a group of diverse and representative individuals. Two more recent deliberation studies report effects on outgroup warmth in addition to policy attitudes and find a reduction in affective polarization among US partisans (Fishkin et al., 2021; Levendusky & Stecula, 2021).

While expressly deliberative interventions rely on participant rationality (see Sanders (1997) for example), other forms of interpersonal contact emphasize mechanisms that are more emotional in nature, like humanization, perspective-taking, and empathy. In particular, intergroup contact interventions foreground and seek to affect sentiment about group membership in ways that deliberative interventions do not. While studies of intergroup contact across diverse social cleavages find that such exchange can work to reduce prejudice (see meta-analyses by Paluck, Green and Green (2019); Pettigrew and Tropp (2006)), negative attitudes towards outgroups may remain firmly entrenched even after contact (Mousa, 2020; Scacco & Warren, 2018), and the success of contact may depend on the interaction between the highlighted cleavage and other cross-cutting identity groups (Paler, Marshall and Atallah, 2020). There is mixed evidence around whether contact with an outgroup member is required to induce perspective-getting, or if prompts to consider an outgroup member's perspective without contact can also shift attitudes (Kalla & Broockman, 2023; Tuller et al., 2015).

In the next section, we describe a novel intervention that has features of both deliberation and intergroup contact, and that combines both informational and emotional components. It innovates on deliberative exercises by structuring deliberation and self-reflection *within* a single partisan group, while outgroup members serve as witnesses. In this way, it foregrounds group membership in the same way that most intergroup contact interventions do. But unlike intergroup contact, the intervention we study involves no collaboration between groups, only within them. Moreover, while the intervention explicitly aims to reduce outgroup animus, it targets *partisan* polarization as the outcome of interest, which is rare among studies of the intergroup contact model.

A New Model: Reciprocal Group Reflection

Braver Angels "Red/Blue" workshops were designed to apply principles from couples therapy to the problem of partisan polarization (Doherty, 2021). While the Braver Angels model shares some similarities with previous interventions, it differs in its emphasis on reciprocal group reflection rather than contact, deliberation, or

[2] Santoro and Broockman (2022) illustrate the necessity of such structured dialogue, as unstructured cross-partisan dialogue on areas of disagreement fails to reduce outgroup animus.

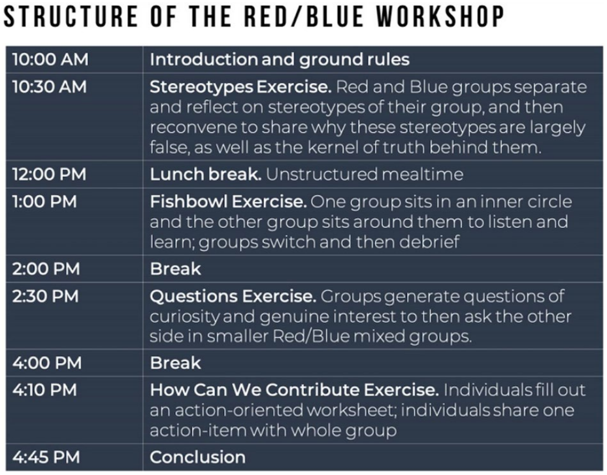
consensus-building. Red/Blue workshops integrate components of previously successful depolarization efforts—such as the provision of information about the policies and perspectives of out-partisans (Ahler & Sood, 2018)—with empathy-building exercises inspired by social psychology (Batson & Ahmad, 2009). The workshops thus combine cognitive and emotional mechanisms, seeking to reduce affective polarization by providing information that counteracts narratives about out-partisans while simultaneously fostering out-party empathy and humanization.

More concretely, Red/Blue workshops are full-day events that engage equal numbers of Republican- and Democratic-leaning participants in a series of structured, moderated exercises, summarized in Fig. 1 and described in more detail in Supplement Appendix A. The workshops involve 5–8 attendees per partisan group and 5–8 observers, depending on the number of participants who sign up. Observers only witness and do not speak during the exercises, but they interact freely with other participants during lunch and breaks. Two Braver Angels volunteers serve as moderators; they go through an online moderator training, consistent across all local Braver Angels "alliances," and attend workshops as observers before they are eligible to lead.

At the beginning of the workshop and prior to the exercises, participants and observers introduce themselves and identify as "Reds" (Republican and Republican-leaning) or "Blues" (Democrat and Democrat-leaning). Moderators lay out the ground rules, mainly that participants (1) are there to understand and express diverse views, not convince anyone to change their mind; (2) speak for themselves; (3) listen actively and participate in the spirit of the activities; and (4) act with respect toward the workshop participants and observers.

How Reciprocal Group Reflection May Reduce Affective Polarization

As we discuss below, we find consistent evidence that this novel model of reciprocal group reflection substantially reduces affective polarization. Because the only contrast we experimentally manipulate is between the treatment and control groups, we



STRUCTURE OF THE RED/BLUE WORKSHOP

10:00 AM	Introduction and ground rules
10:30 AM	**Stereotypes Exercise.** Red and Blue groups separate and reflect on stereotypes of their group, and then reconvene to share why these stereotypes are largely false, as well as the kernel of truth behind them.
12:00 PM	**Lunch break.** Unstructured mealtime
1:00 PM	**Fishbowl Exercise.** One group sits in an inner circle and the other group sits around them to listen and learn; groups switch and then debrief
2:00 PM	Break
2:30 PM	**Questions Exercise.** Groups generate questions of curiosity and genuine interest to then ask the other side in smaller Red/Blue mixed groups.
4:00 PM	Break
4:10 PM	**How Can We Contribute Exercise.** Individuals fill out an action-oriented worksheet; individuals share one action-item with whole group
4:45 PM	Conclusion

Fig. 1 Day-long Red/Blue workshop

cannot fully pin down the mechanisms through which the Red/Blue workshops affect polarization. However, we conducted systematic observations of the workshops and analyzed coded transcripts to identify how unique features of the model may overcome some of the challenges that can otherwise undermine intergroup contact. As a theory-generating exercise, we summarize key intuitions here. In the final section, we discuss how future work might systematically test these propositions.

First, unlike many interpersonal and intergroup contact studies, perspective giving and getting in our intervention is expressly reciprocal. Each ingroup member knows that, just as they receive perspectives from the outgroup, the outgroup also receives perspectives from them. This reciprocity may generate a greater openness to new perspectives, with the knowledge that costly information updating will be mutual. In other words, knowledge of an implied mutual contract in which both sides send and receive perspectives may be important to the willingness of each group to take up new perspectives. That each side takes turns demonstrating some level of vulnerability (e.g., in questioning their own party's ideas or identifying a kernel of truth in stereotypes about them) may further engender empathy and open-mindedness. Importantly, this process does *not* involve deliberation, debate, or attempts at consensus-building or policy compromise, which may be counterproductive in some circumstances (Mutz, 2006), and unlikely to be an effective depolarization strategy where many citizens are not far apart on policy preferences in the first place (Mason, 2015).

Second, we posit that one of the strengths of the intervention—and the reciprocal group reflection model more generally—is its potential to activate informational and emotional mechanisms simultaneously and in a mutually reinforcing way. For example, humanizing the outgroup may build empathy (Andrighetto et al., 2014; Cassese, 2021), and building empathy may in turn facilitate the processing of new information about the outgroup (Cook, 1978; Miller, 2002). This distinguishes reciprocal group reflection from deliberative exercises which emphasize informational mechanisms over emotional ones and generally have the goal of facilitating policy debate and compromise rather than reducing outparty animus.

A third unique characteristic of the Red/Blue model relative to some studies of interpersonal contact and perspective-getting is that outgroup views and opinions are transmitted by a group rather than a single individual. This may help participants generalize learnings from interpersonal contact to members of the broader group—a process that involves what Pettigrew (1998) calls de-categorization, salient categorization, and re-categorization. As we show in our discussion of qualitative data in Sect. 5, workshop activities foster empathy as participants listen to the perspectives of outgroup members, e.g., in the Stereotypes and Fishbowl exercises described in Fig. 1. In the process, participants "de-categorize" or disassociate individual outgroup members from the stereotypes they hold about the outgroup as a whole. By hearing from their ingroup, workshop members also learn about the diversity of values and perspectives within their party, disrupting more positive stereotypes they might hold about co-partisans. The workshop then stimulates "salient categorization" by reminding participants that outgroup members are typical of their group (through, for example, the Fishbowl exercise). Finally, learning about shared values through activities such as the Questions exercise makes re-categorization—adopting an inclusive category that highlights similarities and obscures boundaries—possible.

Research Design

We evaluated the impact of Red/Blue workshops during Spring 2020 on four college campuses in the Northeast, Mid-Atlantic, South, and Midwest. College students are a particularly relevant population for this study, and for programs aimed at reducing polarization more generally. Analysts have expressed particular concern about the heightened risk of polarization among young people in today's political environment (Iyengar et al., 2019). In fact, freshmen entering college in 2017 were the most polarized class in 50 years (Glatter, 2017), and college graduates today are more likely to hold consistently liberal or conservative positions compared to previous cohorts and those with a high school education today (Pew, 2016). At the same time, this population represents the future of partisan politics in the country, with possibly the greatest potential to become "norm entrepreneurs" (Finnemore & Sikkink, 1998) whose commitment to cross-partisan dialogue is critical to the broader success of depolarization initiatives (Santos et al., 2022). While we cannot say for certain whether our results would generalize to the broader population, in Supplement Appendix C we show that our student sample is broadly comparable to the American public on key baseline characteristics and attitudes, which suggests that our findings may be applicable beyond college campuses[3].

Recruitment, Randomization, and Survey Administration

To recruit study participants, we worked with two student organizers on each campus, one from the College Democrats and one from the College Republicans, mimicking the process of the larger Braver Angels organization in which local Red and Blue organizers are tasked with recruiting participants for each workshop. We also solicited student expressions of interest via departmental and student group listservs[4]. All undergraduate US citizens were eligible to apply. Interested students were sent additional information about the study and asked to complete an online survey that included demographic information, as well as baseline measures of partisanship, polarization, and other political attitudes.

We recruited approximately 40 students at each university to participate in the study and complete the baseline using a self-administered Qualtrics survey sent via email approximately two weeks prior to the workshops held in February and March 2020. Participants included a total of 165 subjects, including 116 "Blues" and 49 "Reds"[5]. Of those who completed the baseline, 59 were randomly assigned to partici-

[3] Given the Red Blue Workshop's voluntary nature, our sample consists of individuals with at least some interest in participating in this sort of event. Even if our findings generalize only to those with some interest in cross-partisan interaction, this is the population of interest if the goal is realistic progress in curtailing growing polarization in the US.

[4] Though not all participants were members of the College Democrats or Republicans, we expect those who were to be more politically engaged and have more entrenched political views. So we might expect effects among this population to be smaller than among a more politically malleable one.

[5] Braver Angels is a non-partisan organization with conservative and liberal leadership, and aims to recruit equally across parties. Consistent with experiences of Braver Angels nationally, however, recruitment on campuses attracted more liberal-leaning students even at the more conservative-leaning campuses in our

pate in the workshop as either participants or observers; the remainder were assigned to control. Random assignment to the workshop was blocked by campus, partisan ID, and baseline affective polarization. Midline surveys, self-administered using Qualtrics in the same manner as the baseline, were sent via email to participants one to two weeks after the workshop, and the endline was sent roughly six months after that in October 2020. Members of the control group were invited to complete all follow-up surveys at the same time as the treatment group.

Per Braver Angels policy, each workshop had roughly the same number of participants from each party. At two of the four universities, all Red-leaning students who completed the baseline were assigned to treatment, given the uneven number of Republicans and Democrats recruited on those campuses and the minimum number of Red-leaning participants required to hold a workshop. Because we include campus fixed effects in our analyses, these Red-leaning students do not contribute to our treatment effect estimates. Participant and observer roles within the treatment group were randomly assigned on the day of the workshop, stratifying by partisanship[6]. We consider the ethical implications of our study and describe our risk assessments and the measures we took to minimize risks and maximize benefits for participants in Supplement Appendix B.

Power

We report power calculations based on simulations using baseline values of our dependent variables in Supplement Appendix E. While our sample size is modest, our simulations suggest that we are nonetheless powered to detect effects of approximately 0.21 standard deviations on our explicit measure of affective polarization and approximately 0.45 standard deviations on our implicit measure. Comparing these effect sizes to different standards in the literature, we show in Supplement Appendix E that the former is very much in line with effects on explicit outcomes in previous studies, and the latter is reasonable for a relatively intensive intervention.

Moreover, as we will see, our intention-to-treat (ITT) estimates are consistently significant across our explicit, implicit, and behavioral measures. They are also substantively similar (albeit underpowered) when we analyze each campus separately. This consistency suggests that our results are unlikely to be artifacts of Type I errors induced by a small sample size[7]. Furthermore, we find consistently significant effects on our primary outcomes—those that the intervention was most explicitly designed

sample. While we lack direct evidence to explain the sources of this heterogeneity, prior research points to differences in personality traits such as greater openness to new experiences among liberals (Gerber et al., 2011), greater perceptual flexibility or tolerance of ambiguity among liberals (Jost, 2017), and even the differential structural functions of parties (Grossman & Hopkins, 2015).

[6] In our pre-analysis plan (PAP) we proposed to test the effects of the workshops on participants and observers separately. Because we had a small number of observers, we collapse these two categories in our analysis.

[7] If these significant results were simply due to random noise, then we would not expect to find consistent effects across different outcome measures or campuses. If solely due to random noise, it would be highly unlikely for an effect to show up in the same direction across different measures or campuses; the estimated effects should randomly vary around zero, so there would be an equal probability of effects or non-effects in different directions.

to affect—but not on a number of other outcomes that we measured. If our results were merely Type I errors attributable to a small sample size, then we would expect the results to be more idiosyncratic; the fact that all the significant effects we estimate are concentrated among the small number of outcomes that are the most closely related to the goals of the intervention—and which we pre-specified as being primary—suggests they are not simply artifacts of Type I errors.

Measurement

Our primary outcome of interest is affective polarization, which we measure both explicitly and implicitly using our surveys. We measure an additional related primary outcome behaviorally: support for depolarization[8]. We also use our surveys to construct indices corresponding to the pre-registered mechanisms described above, and to operationalize four secondary outcomes that were less central to the goals of the workshops: support for pro-democratic politicians and actions, support for civil discourse, capacity for civil discourse, and ad hominem attributions. We focus here on our primary outcomes, and report (mostly null) treatment effects on our pre-specified mechanisms in Supplement Appendix J, and on our secondary outcomes in Supplement Appendix P.

We operationalize *explicit* affective polarization using a series of direct questions administered in all three waves of the survey. These include (1) a “feeling thermometer” capturing the difference between respondents’ feelings towards the in-party and out-party[9]; (2) the difference between respondents’ trust in the in-party and out-party[10]; three-point Likert scales capturing how comfortable respondents would feel having out-partisans as (3) close personal friends and (4) neighbors, and (5) as spouses of their best friend; and (6) a dummy for respondents who believe the out-party represents a “serious threat” to the country. To construct our explicit polarization index, we first standardize and aggregate our three measures of comfort with out- partisans (measures 3, 4, and 5 above) into a single index, leaving us with four measures of affective polarization (measures 1, 2, and 6 above, plus the out-partisan comfort index). We then aggregate standardized versions of these four measures into a single index. Finally, we re-standardize the midline and endline values of this index to its baseline value[11].

[8] In our PAP we proposed to measure two additional behavioral outcomes: an indicator for initiating conversations with out-partisans, and another indicator for participating in additional depolarization interventions between our midline and endline surveys. The first of these measures is based on survey self-reports; as a result, in retrospect we do not believe it can be meaningfully interpreted as a behavioral measure. The second of these measures we were unable to implement due to the cancellation of additional depolarization interventions in response to the COVID-19 pandemic. Nonetheless, we report treatment effects on the first outcome and further discuss the second in Supplement Appendix I.

[9] We ask respondents to rate how favorably they feel towards the in-party and out-party on a scale of 0 to 100. The feeling thermometer is the difference between these two numbers.

[10] We ask respondents to report how often they believe they can trust the in-party and out-party to “do what is right for the country” on a scale of 1 to 5, where 1 indicates almost never and 5 indicates almost always. We take the difference between these two numbers.

[11] We did not pre-specify that we would standardize the midline and endline values of our outcomes to their baseline values. This represents a minor deviation that allows us to our ITT estimates *relative to baseline*

We measure *implicit* affective polarization using an Implicit Association Test (IAT) administered in all three waves of the survey[12]. Respondents were shown a series of images typically associated with either Democrats or Republicans (e.g. a donkey or an elephant, respectively). On either side of each image they were shown the names of the two parties paired with one of two evaluative words (e.g. “Democrat or bad” or “Republican or good”). They were then asked to associate each image with the corresponding party as quickly as possible. The assumption underlying the test is that respondents experience cognitive dissonance when asked to pair the name of the out-party (in-party) with a positive (negative) evaluative word, causing their response times to slow down (Greenwald et al., 1998). We measure implicit affective polarization by taking the standardized difference in in-party and out-party response times, then re-standardizing the midline and endline differences to their baseline values.

We measure support for depolarization behaviorally using a simple donation prompt administered at midline and endline. Before beginning the midline, respondents were informed that they would receive a $10 Amazon gift card as compensation for completing the survey[13]. At the end of the survey, respondents were given the option of donating a portion of their compensation to Bridge the Divide, a separate NGO whose goal is to reduce polarization among American youths. Respondents could donate in increments of $5. We interpret the amount donated as a measure of support for depolarization, i.e. respondents’ willingness to take a more “costly” action to reduce partisan polarization and promote a civic culture conducive to democracy. About a quarter of respondents chose to donate some of their compensation at midline and also at endline. Since we did not measure this outcome at baseline, we standardize endline values to midline values only.

By employing three different outcome measures, we guard against the possibility that our results are driven by the particularities of any one indicator. The survey questions that comprise our index of explicit affective polarization are widely used in the field, but are potentially subject to demand effects. The IAT protocol used for our implicit measure of affective polarization is designed to overcome social desirability and other biases and measure underlying prejudices; even critics characterize the IAT as a valid measure of political attitudes (Schimmack, 2021), and large-scale meta-analyses and replications suggest that implicit attitudes are pervasive, predictive of behavior, and highly correlated with explicit bias (Greenwald et al., 2006). The behavioral measure, meanwhile, leverages real, monetary outcomes to make misreporting more costly, even if donation decisions are a noisy proxy for political behavior[14]. If the intervention affects each of these measures in similar ways, this should increase our confidence in the robustness of our estimated effects. As an exploratory exercise, in Supplement Appendix M we aggregate across our various measures of

but does not change the corresponding t-statistics or p-values.

[12] There does not seem to be any learning effects over time, as our control group respondents do not exhibit a decline in polarization on the IAT.

[13] We increased this amount to $20 at endline to mitigate attrition.

[14] Although young people make up a small percentage of political donations in the US (Huges, 2017), the recent rise in small dollar donations has likely helped to normalize relatively modest donations among youths. News outlets also reported a notable uptick in $5-$10 donations from college students following the Summer 2020 protests for racial and criminal justice reform (Mier, 2020).

affective polarization and support for depolarization using the Average Effect Size (AES) estimator proposed by Clingingsmith et al. (2009) and Kling et al. (2004)[15].

Empirical Specification

Our pre-analysis plan specified a pooled analysis of midline and endline outcomes. We report these results in Supplement Appendix L, but focus here on a more informative way to model the data: a difference-in-differences analysis that allows us to estimate marginal effects at midline and endline separately while also adjusting for any baseline differences between the treatment and control groups. This specification is also advantageous because it directly captures decay in the magnitude of our intention-to-treat (ITT) estimates over time, which is obscured by the pooled analysis. This is especially important given the relatively long lag between the workshops and the endline, and the precipitous unforeseen changes in the political landscape that occurred in the interim (including a politicized pandemic, nationwide protests for racial justice, and a highly divisive presidential election). We summarize this and all other deviations from our PAP in Supplement Appendix H.

Formally, we estimate

$$Y_{ij} = \alpha + \beta_1 Treatment_i + \beta_2 Wave_j + \beta_3 (Treatment_i \times Wave_j) + \beta_4 Block_b + \varepsilon_{ij}$$

where $Treatment_i$ denotes treatment assignment, $Wave_j$ denotes the survey round (baseline, midline, or endline), and $Block_b$ indexes 16 blocks based on campus, party, and baseline affective polarization (above or below the median for a particular campus and party). We use this difference-in-differences estimator to compute marginal effects at midline and endline. Standard errors are clustered by respondent.

Results

Figure 2 summarizes the ITT of the intervention on our three main outcomes at midline and endline, using the difference-in-differences estimator described above. All outcome measures are standardized to facilitate comparison; for purposes of consistency, the behavioral measure is reverse-coded such that negative treatment effects imply larger donations to Bridge the Divide. All estimates include block fixed effects. A tabular version of the regression results used to compute marginal effects for Fig. 2 is in Supplement Appendix R.

We observe a statistically significant negative ITT on each of our three main outcomes at midline (March 2020); the effect is largest for our behavioral measure, smaller for our implicit measure, and smallest (and not quite statistically significant at the 5% level, $p = 0.057$) for the explicit measure. These effects decay over time for the explicit and implicit measures of affective polarization, losing statistical significance at endline (October 2020), though the ITTs remain negative[16]. The effect on

[15] This latter analysis was not pre-specified.

[16] We present the pre-specified pooled results in Figure L.1, where the ITT on the explicit measure is no longer statistically significant at conventional levels (p = 0.270). The ITTs on the implicit and behavioral

our behavioral measure persists, remaining large and statistically significant even at endline. This is reassuring, as the behavioral measure is arguably the one that is least susceptible to experimenter demand effects, and the midline and endline were fielded six months apart.

Our index of explicit affective polarization incorporates multiple measures of perceptions of the in-party and out-party. As scholars have shown, however, the growth in affective polarization in recent years has been driven by increasing hostility towards the out-party, rather than increasing affinity for the in-party (Iyengar & Krupenkin, 2018). Our results are consistent with this trend. Figure 3 disaggregates the treatment effect on affective polarization by feelings and trust towards the in-party and out-party. We find that the treatment effects at midline in Fig. 2 are driven by a reduction in out-party hostility rather than a reduction in in-party affinity.

To put these estimates in perspective, between 1978 and 2020, the average out-party thermometer rating in the nationally representative American National Election Studies (ANES) survey dropped from roughly 52.61 to 19.56—a decline of approximately 33.05 points. In our midline, the average out-party thermometer rating among treatment group participants was 6.78 points higher than the average rating among control group participants. If a treatment effect of this magnitude were extrapolated to the ANES sample, it would reverse approximately one-fifth of the decline in out-party "warmth" observed over more than four decades. Similarly, between 1978 and 2020, the difference between in-party and out-party ratings in the ANES feeling thermometer grew by roughly 31.30 points. In our midline, the average feeling thermometer difference among treatment group participants was 7.43 points smaller than

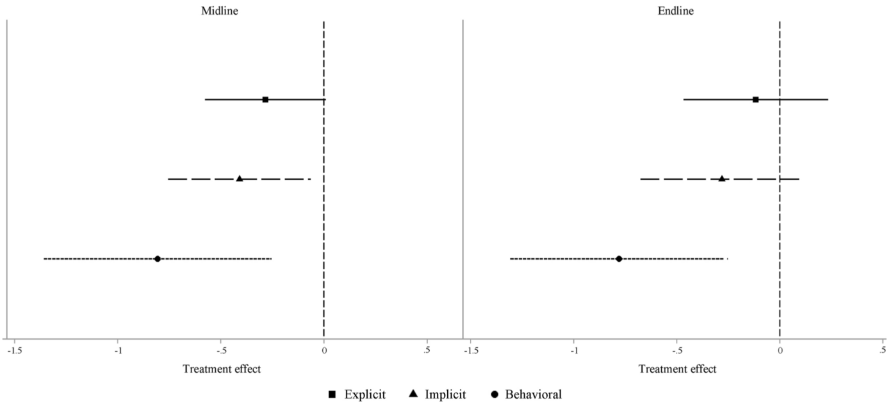


Fig. 2 Treatment effects on affective polarization and support for depolarization. *Note:* Estimates are standardized and include block fixed effects. Treatment assignment was blocked on (1) campus, (2) party, and (3) baseline affective polarization. Standard errors used to calculate 95% confidence intervals are clustered by respondent

measures remain strongly statistically significant regardless of specification. Figure L.2 reports treatment effects estimated separately at each round for midline and endline (i.e. by running two separate regressions, rather than computing marginal effects from the difference-in-differences estimator). The results largely confirm the conclusions of Fig.2 although the ITT on the explicit measure is somewhat weaker.

the average difference among control group participants. If a treatment effect of this magnitude were extrapolated to the ANES sample, it would reverse nearly one-quarter of the increase in affective polarization observed over more than four decades.

An intervention of this sort is likely to be most successful if it reduces affective polarization among both parties. While our sample size is too small to conduct a well-powered test of the difference in effect sizes between Republican and Democratic participants, Fig. 4 shows that the treatment effects on Democrats and Republicans (as well as on Republican- and Democratic-leaning independents) look strikingly similar, with no indication of significant differences between them. These results suggest that the intervention reduced affective polarization among Republicans and Democrats alike. Results from additional pre-specified analyses of heterogeneous treatment effects can be found in Supplement Appendix K.

Given substantial attrition across the three waves of the survey (29%), one might be concerned that the apparent treatment effects on our main outcome measures are artifacts of systematic differences in the composition of the sample (akin to a selection effect) induced by differential attrition across the treated and control conditions. We probe this possibility in Supplement Appendix Q. We find that the primary predictor of attrition is treatment status, with treated respondents significantly more likely to respond to the midline and endline surveys (22 and 11 percentage points, respectively) than those in control. As Supplement Table Q.1 shows, differences between attriters and non-attriters are otherwise generally small and not statistically significant except for a few variables in the endline.

Building on this analysis, in Fig. 5 we present attrition-adjusted treatment effect estimates using an inverse probability weighting (IPW) approach, which recovers the ATE under the assumption that attrition is independent of potential outcomes condi-

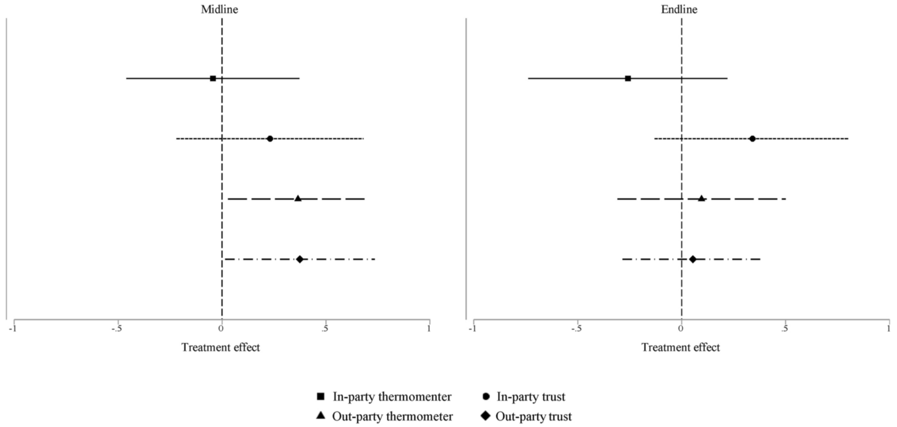


Fig. 3 Treatment effects on affective polarization, in-party vs. out-party. *Note:* Estimates are standardized and include block fixed effects. Treatment assignment was blocked on (1) campus, (2) party, and (3) baseline affective polarization. Standard errors used to calculate 95% confidence intervals are clustered by respondent

tional on covariates (Gerber & Green, 2012, 221)[17]. Our treatment effect estimates using this procedure remain substantively similar in magnitude for all three of our main outcome measures, though the IPW-adjusted effect on implicit polarization at endline becomes stronger and retains significance at the $p < 0.05$ level.

One might also wonder about the possibility of experimenter demand effects driving the results, which would occur if treated individuals reported lower levels of polarization due to internalized expectations that they *should* appear less polarized as a result of participating in the study. But this seems unlikely to explain our findings. Both treatment and control group respondents were aware of the purpose of the workshop and the study, so the reporting of socially desirable attitudes and behaviors around polarization could easily apply to both groups. Yet despite reporting similar levels of polarization at baseline, the groups diverge post-treatment, with respondents in the treatment condition showing reduced polarization relative to those in control.

Moreover, as we show in Supplement Appendix J, we find no effect on related outcomes (such as stereotyping or humanization of the outgroup) that should be similarly susceptible to demand effects; if treated respondents were parroting back what they thought we wanted to hear, we should have picked up effects on these outcomes as well. In fact, our most persistent treatment effects are on implicit attitudes and the behavioral measure—the outcomes that should be *least* likely to be influenced by experimenter demand—which is the opposite of what we would expect if demand effects were indeed driving the results. Finally, state-of-the-art evidence on the prevalence and magnitude of demand effects in social science experiments demonstrates that they appear to be quite limited as a general rule (Mummolo & Peterson, 2019). The data suggests that our own study is in line with this overall trend.

Mechanisms

We find that recriprocal group reflection reduced affective polarization, at least in the short term. While our PAP outlined seven potential mechanisms that might explain these results, as we show in Supplement Appendix J, we find little to no evidence for any of these mechanisms in the survey data. Because we *do* find treatment effects on affective polarization, the lack of quantitative evidence for our pre-specified mechanisms suggests either that the effects are being mediated by mechanisms we did not measure, or that our measurement strategy did not successfully operationalize the underlying constructs. In an effort to uncover the mechanism(s) at work, we exploit the rich qualitative data contained within transcripts of the workshops. It is from this analysis that we inductively identified the mechanisms outlined in our theoretical framework. We argue that the workshops reduced affective polarization by simultaneously increasing empathy for the outgroup (an emotional mechanism) and induc-

[17] We describe the construction of these inverse probability weights in Supplement Appendix Q. These analyses were not pre-specified. However, Gomila and Clark (2022, 148) note that when treatment assignment causes missingness (as in our case), "it is more likely for missingness to be conditional on a set of covariates," making this a seemingly appropriate case for the applicaton if IPW.

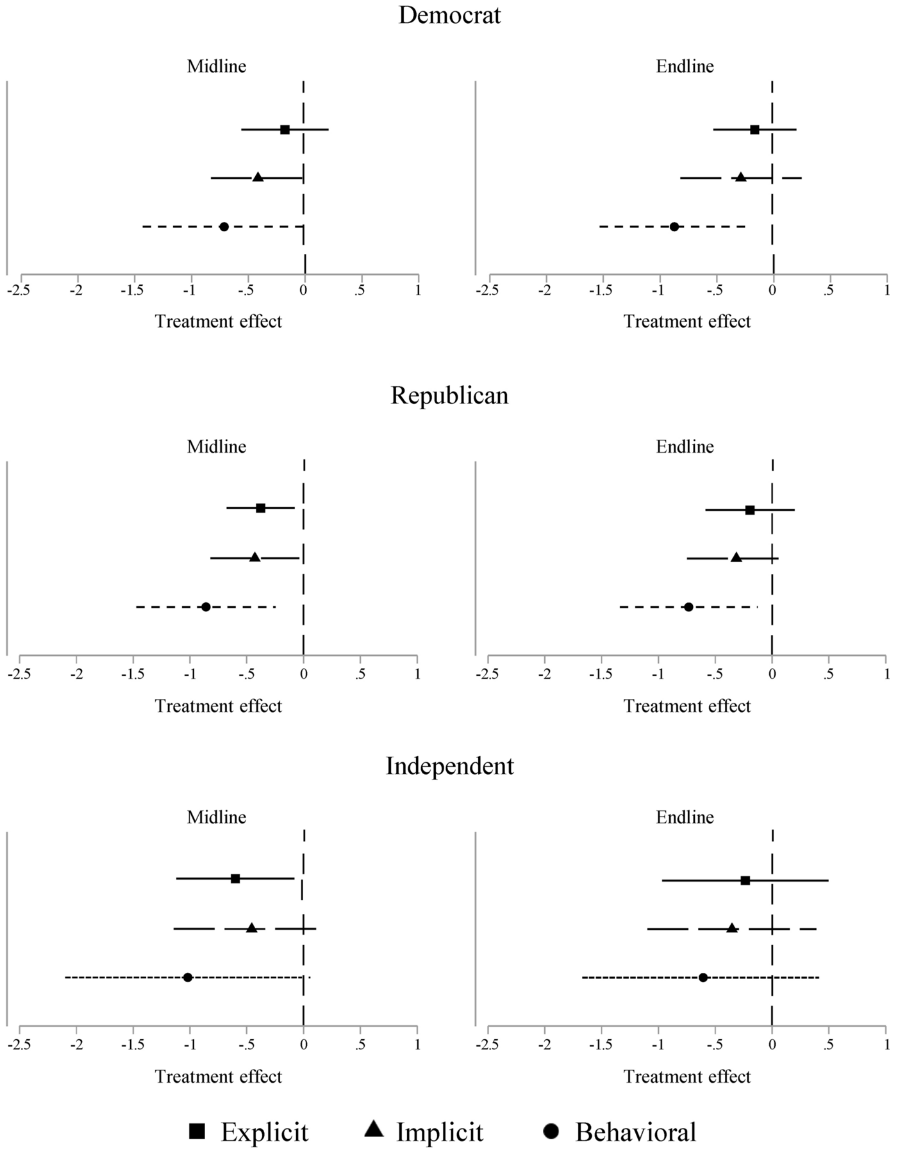


Fig. 4 Treatment effects on affective polarization and support for depolarization by party. *Note:* Estimates are standardized and include block fixed effects. Treatment assignment was blocked on (1) campus, (2) party, and (3) baseline affective polarization. Standard errors used to calculate 95% confidence intervals are clustered by respondent

ing learning about cross-partisan commonalities and intra-partisan heterogeneity (a cognitive mechanism).

We audio-recorded and transcribed all workshop activities and assigned pseudonyms to participants to protect their privacy. We use this data—akin to four focus groups—in two ways. First, a team of research assistants coded the transcripts in rela-

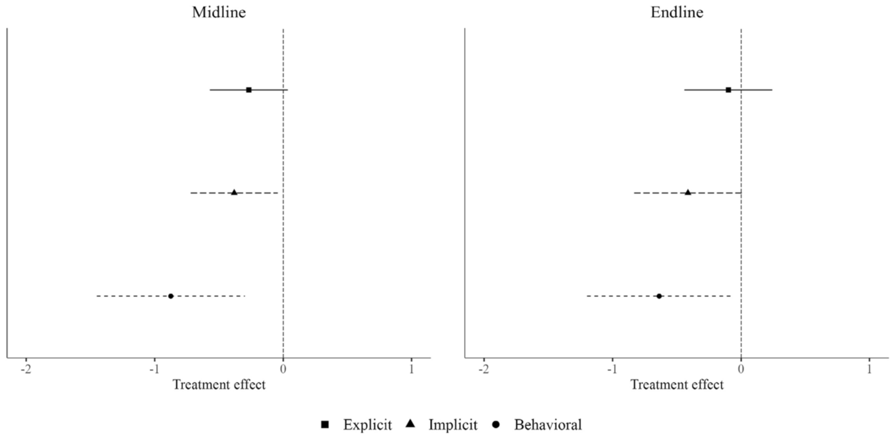


Fig. 5 Treatment effects on affective polarization using IPW for attrition. *Note:* IPW-adjusted treatment effect estimates are standardized and include block fixed effects. Variables included in the IPW approach are based on Supplement Appendix Table Q.1

tion to the seven pre-specified mechanisms, using a codebook and manual available in the Supplementary Appendix. In particular, we classified units of data (phrases or full participant interventions) based on the type of mechanism they exemplified and their directionality (i.e. whether they signaled polarization or depolarization), identifying a total of 722 depolarizing remarks (see frequencies in Fig. 6). Second, because depolarization during the workshops is an interactive process, we also read the transcripts and analyzed conversations as a whole to capture group interactions, illuminate patterns behind the frequencies in Fig. 6, and identify alternative depolarization pathways we had not anticipated or pre-specified. Together, these analyses help illustrate the mechanisms behind the reciprocal group reflection model described above.

We identify how both emotional and informational components combine in the intervention to allow individuals to generalize from their experience in the workshop to broader ideas about their ingroup and outgroup. Comments and conversations that (1) recognize ingroup heterogeneity, (2) show empathy for the outgroup, and (3) identify commonalities across groups are particularly salient throughout the workshops. They illuminate how participants *decategorize* in- and outgroup individuals from party stereotypes and learn to understand the outgroup point of view; experience *salient categorization* by discussing shared group values and priorities; and undergo a process of *re-categorization* by recognizing similarities across groups (beyond just the individuals in the workshops). Reciprocity between groups and group-based perspective-getting—two key features of the Braver Angels model—appear to have created an emotional framework that facilitated the cognitive processes described below.

Decategorization

Decategorization happened in two ways. First, the workshops decreased ingroup identity salience by inducing participants to perceive their own party to be more heterogeneous than previously assumed. Importantly, the recognition of ingroup

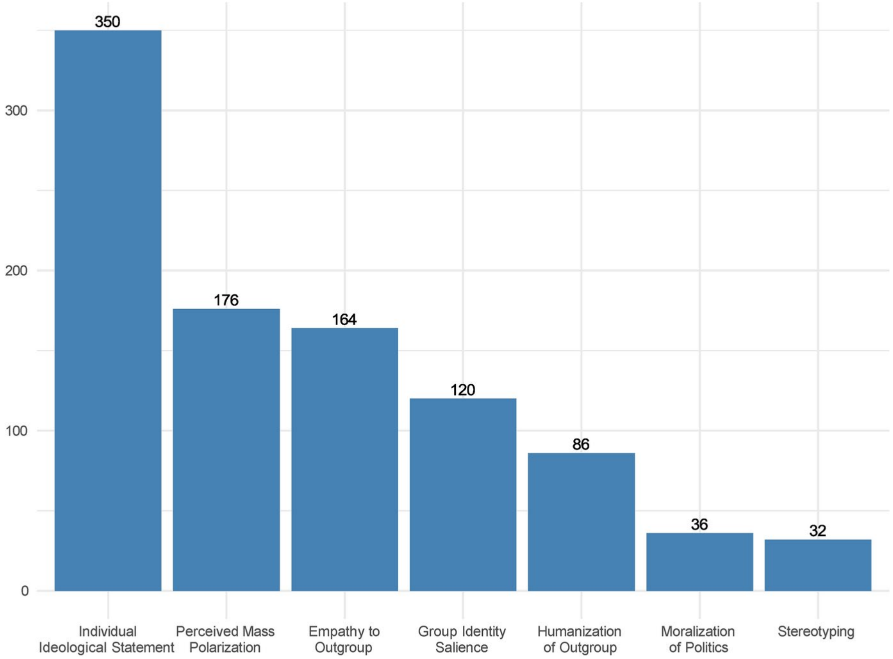


Fig. 6 Frequency of discussion topics. *Note:* Number of times each pre-specified mechanism appeared across all workshops. Mechanisms were coded with a negative or positive valence in mind (i.e. whether they indicate increased or decreased polarization). One exception is individual ideological statements in which we coded any statement related to a participant's ideological position, and were difficult to categorize as evidence of polarization or depolarization. For the other mechanisms, we only report codes that go in the depolarizing direction as we are interested in mechanisms of depolarization

diversity occurred in front of the outgroup, increasing mutual vulnerability. Comments signaling decreased ingroup identity salience were the third most common mechanism identified in the workshop transcripts (120 instances); comments related to ingroup heterogeneity were especially frequent (111 instances). Ingroup heterogeneity was explicitly acknowledged by participants, as in the following quote from the Questions exercise:

> Abby (Blue): So I am a pro-life Democrat... The thing that comes up with that, it sort of assumes you have to agree with every single thing about your party. And I think that is completely unrealistic. And if anything it limits American democracy by not allowing people to be within a party and not believe totally [in everything it stands for].

Such language was also common in discussions about ingroup "weaknesses" in the Stereotype and Fishbowl exercises. As illustrated by the conversation below discussing the "racist" stereotype, students often framed the "kernel of truth" by questioning their ingroup homogeneity. Separat- ing themselves (and others) from more extreme (stereotypical) voices in the party, they often acknowledged that "some, but not all" in their party might fulfill the stereotype being discussed:

> Jed (Red): Guaranteed there is a subset that are completely racist. Because I've heard it myself. But I feel that's false for the vast majority...

> Pauline (Red): The loudest people don't represent the majority of the people.

De-categorization or disassociation also happened with respect to the outgroup. Empathy for the outgroup—that is, the ability to "imagine how another person thinks and feels given their situation, or an imagine-other perspective" (Batson & Ahmad, 2009)—was the second most commonly coded mechanism in the transcripts. Participants frequently expressed an increased willingness and capacity to see things from the point of view of out-partisans. Maya (Blue), for example, shared that to "sit down and understand the thought process and try and empathize with the other point of view…[gave her] some strategies to do that outside of [the workshop] and navigate that." Interested in the topic of religion, Tony (Red) appreciated "the very individual responses" that Blues gave, and was pleased that one can "actually have meaningful interactions…when you take the time to sit down with people to listen and to have a respectful conversation over the course of a couple hours… Once you have that transparency and discussion about what people believe, you try to find commonalities, and don't just stereotype."

The reciprocal nature of the exercises appears to have generated greater emotional openness to receiving this new information. Some activities—especially the Stereotypes and Fishbowl exercises— are designed in part to use reciprocity to create a sense of shared vulnerability. Participants are asked to identify the "kernel of truth" underlying negative stereotypes about their ingroup, and are encouraged to voice doubts about their own party's values and policies, all while being observed by members of the outgroup. As suggested by the quote below, by inducing participants to let their guard down—and to listen as both co-partisans and out-partisans let their guard down as well—these exercises helped establish the mutual empathy that facilitates information updating (i.e. re-categorization):

> Ray (Blue): "I also found there to be, you know, lots of really good introspection on both sides. And I felt really validated by a lot of the self-reflection points during the Fishbowl especially."

> Yvonne (Blue): "I definitively was surprised by the amount of self-reflection and political criticism... It made me actually self-reflect and realize that maybe I should be more critical of my party."

Salient Categorization

The disassociation mechanisms outlined above occurred alongside a process of salient categorization. Participants' partisan identities are reinforced through the group-based exercises, and participants are asked throughout the day to speak as a supporter of their party. This process is perhaps clearest in the Fishbowl exercise, when participants are asked to discuss reasons their side's values and policies are good for the country, as well as reservations they have about their own party. This setup allows participants to connect their individual preferences (and themselves) with values and policies espoused by "their" group. In one conversation, Mollie (Blue) stated, "I think it is important to acknowledge no matter what party is dominant in political offices, that our country is very prone to institutionalized discrimination. And that's something that *our party* aims to resolve." Similarly, Howard (Blue) explained that he thinks the values and policies of his party are good for the country because "*our side* recognizes everyone as they come from different backgrounds."

The debrief at the end of the Fishbowl exercise demonstrates how it helped participants think of these conversations as credible—given by, and representative of, the outgroup. Bert (Blue), for example, learned that "*they* [Reds] are at a crossroads, especially when it comes to the whole nationalist versus globalist stand." He now believes that "*the Red side* understands that *they are* at a crossroads where *they're* going to have to decide one way or the other how involved *they're* going to be in the world." The group-based structure of the intervention appears to have facilitated this process of salient categorization, which is key to making inferences from interpersonal interactions about a group at large. The intervention emphasized "Red" and "Blue" labels, organizing exercises and ingroup and outgroup conversations in terms of "sides," thus reinforcing the salience of partisan identity, while simultaneously subverting stereotypes through the process of de-categorization described above. Structuring the exercises around partisan identity helped to signal the credibility of workshop participants' perspectives as representative of views within their party writ large.

Re-Categorization

Finally, we observe re-categorization when participants identify commonalities across groups. As depicted in Fig. 6, the most commonly coded mechanism in the qualitative data was a reduction in perceived mass ideological polarization—in particular, the identification of commonalities across groups. Participants often expressed that the workshops helped them understand that they share many more values and goals with out-partisans than they initially realized, even if they disagree on specific policies. For example, Eric (Red) noted that he began "to realize that we all want the United States to do well…and the difference is fundamentally just the means. I realize we really have some ways that we could work together." Similarly, Taylor (Blue) remarked that "once [we] had found that common ground, then we kind of went off into, like I disagree on this, but with regards to certain family values, immediately, both groups. think that's a good idea for kids to be raised in a loving household. And that's where the disagreement happened: on how to implement it. But the…moral foundations are more similar than I had previously thought."

In sum, the transcripts reveal how the intervention activated both cognitive and emotional mechanisms among participants. Comments and conversations during the workshops suggest participants learned that the outgroup is less extreme or more diverse in their views than initially thought; that the ingroup is also more diverse; and that there are more commonalities across groups than participants originally believed. These informational updates occurred against the backdrop of structured, empathy-inducing exercises grounded in a model of reciprocal group reflection.

Discussion

In this study we describe and test a new model for reducing partisan polarization—reciprocal group reflection—based on insights from marital counseling. In an experimental evaluation of Braver Angels' "Red/Blue" workshops, we find statistically significant reductions in explicit and implicit measures of affective polarization, and a statistically significant increase in participants' behavioral support for depolarization. While the effect on our pre-specified explicit measure (difference between in-party and out-party warmth) just misses the 5% significance threshold (p = 0.057), the effect on the now more customary measure of out-party warmth (Hartman et al., 2022; Voelkel et al., 2024) is significant at that threshold. The negative effect on explicit affective polarization is driven in particular by reduced out-party hostility rather than reduced in-party affinity. This is important given evidence suggesting that affective polarization in the US is driven primarily by increasing hostility towards out-partisans (Iyengar & Krupenkin, 2018). These effects appear to be consistent across Republican and Democratic participants.

Effects on the behavioral measure, while more noisily estimated, persist at endline; effects on the implicit and explicit measures remain negative, but lose statistical significance. The weakening of the attitudinal effects could be due to actual decay in the impact of the workshops on affective polarization, or to the radically changed environment between midline and endline. This combination of attitudinal and behavioral results is also consistent with research suggesting that behaviors may be more malleable than attitudes and beliefs (Mousa, 2020; Paluck et al., 2021; Scacco & Warren, 2018).

The exercise in reciprocal group reflection did at least as well as other studied intergroup contact interventions to reduce outgroup prejudice, extending this existing literature on racial and ethnic prejudice to partisan identities. Pooling across intergroup contact studies, Paluck, Green and Green (2019) find that the effect of contact on prejudice is about a third of a standard deviation relative to the control group mean. Our study increases outparty warmth and outparty trust by a very similar amount at midline, about 0.37 standard deviations[18].

Our qualitative data suggest a plausible theory of change: that informational and emotional mechanisms interact to achieve depolarization. Based on systematically coded transcripts of the workshops, we find that the most frequently observed mechanisms include both informational (e.g. perceived mass polarization) and emotional (e.g. empathy towards the outgroup) components. Participants appear to have assimilated new information: they reported learning that the opposing party is less extreme and that their own party is more heterogeneous than they believed (de- categorization); that these lessons about individuals are applicable to the larger group (salient categorization); and that both parties share common values (re-categorization)[19].

[18] It is harder to directly compare the effectiveness of the Red/Blue model to deliberation interventions, as the outcome of interest in studies of deliberation is almost always policy attitudes rather than outgroup prejudice (Theuwis, van Ham and Jacobs, 2021).

[19] This is often what researchers and practitioners hope to achieve with mixed group discussions or contact with the outgroup—both of which have an ambiguous record of success in the literature (Paluck, 2010; Paluck, Green and Green, 2019).

But participants also appear to have changed the way they *feel* about the outgroup, becoming more empathetic and more willing to humanize out-partisans.

We posit that the unique characteristics of the intervention as an exercise in reciprocal group reflection were particularly conducive to depolarization: observing self-reflection in others can transmit new information *and* generate empathy for the out-group. To help generalize lessons from this workshop to other interventions, we would ideally like to know how each of its characteristics drives its impact. Future research could test the value of reciprocity in a much simpler non-group setting by, for example, having pairs of outgroup members listen to each other's perspective, with the knowledge it will be reciprocal, and compare that to a similar unpaired perspective-getting exercise. We would then want to understand whether a reciprocal perspective-getting exercise at the individual level has a similar effect on salient categorization—i.e. applying inferences about a group member to an entire group—as a reciprocal perspective-getting exercise at the group level.

As to why the mechanisms identified in the qualitative data are not apparent in the survey data presented in Supplement Appendix J, we believe this discrepancy is likely due to a mix of (1) unanticipated mediators and (2) mismeasurement. As an example of the former, our survey measure of perceived mass ideological polarization captures the extent to which people believe the public is polarized around certain issues. But the qualitative data show that participants did not necessarily update their priors about issue polarization; instead, they updated about the commonality of values across groups, as well as the heterogeneity of the ingroup—neither of which is captured by our survey data. As for mismeasurement, the fact that we see substantial empathy expressed toward the outgroup during these workshops may indicate that our quantitative measure corresponding to self-reported willingness to take the perspective of an outparty member is insufficiently sensitive to changes in empathic tendencies among participants.

Finally and more speculatively, there are some aspects of this particular intervention that we believe merit further examination. The moderators' adherence to "even-handedness" may have allowed misinformation or extreme views to go unchecked, potentially altering participants' perceptions of what is true or normatively acceptable. This was not something our study was designed to analyze, but we believe it deserves further consideration. We also note that depolarization in the past has sometimes come at the expense of the rights of, and justice for, marginalized minority groups. Levitsky and Ziblatt (2018) describe this as a bargain between elites—whereby lawmakers implicitly or explicitly agree to abandon racial justice in order to placate colleagues who hold.extreme (racist) views—but it is possible that we might observe similar dynamics among workshop participants, especially given the emphasis on harmony and reconciliation. On the other hand, there is accumulating evidence that non-threatening conversations can reduce intolerance (Kalla & Broockman, 2020)—gains that could be lost with a more confrontational approach. This, too, merits further consideration. All told, we see reason to be optimistic about the depolarizing effects of workshops that encourage civil discussion and perspective taking between groups with seemingly irreconcilable views.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11109-024-09993-z.

Funding The funding provided by the Scowcroft Institute of International Affairs at the Bush School of Government and Public Service at Texas A&M University, Institute for Democratic Engagement and Accountability (IDEA) at Ohio State University, Swearer Center and Department of Political Science at Brown University, Department of Political Science at University of Pittsburgh.

Data Availability Data and replication files are available on Harvard Dataverse at https://doi.org/10.7910/DVN/ZDFX1V.

References

Ahler, D. J., & Sood, G. (2018). The Parties in Our Heads: Misperceptions About Party Composition and Their Consequences. *Journal of Politics, 80*(3), 964–981.

Allport, Gordon. 1954. *The Nature of Prejudice*. 2nd ed. Cambridge, MA: The Beacon Press. Andrighetto, Luca, Cristina Baldissarri, Sara Lattanzio, Steve Loughnan and Chiara Volpato. 2014. "Human-itarian Aid? Two Forms of Dehumanization and Willingness to Help after Natural Disasters." *British Journal of Social Psychology* 53:573–584.

Andrighetto, L., Cristina, B., Sara, L., Steve, L., & Chiara, V. (2014). Humanitarian aid? Two forms of dehumanization and willingness to help after natural disasters. *British Journal of Social Psychology, 53*(3), 573–584.

Batson, D. C., & Ahmad, N. Y. (2009). Using Empathy to Improve Intergroup Attitudes and Relations. *Social Issues and Policy Review, 3*(1), 141–177.

Broockman, David E., Joshua L. Kalla and Sean J. Westwood. 2020. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not." OSF Preprint. https://doi.org/10.31219/osf.io/9btsq

Cassese, E. C. (2021). Partisan dehumanization in American politics. *Political Behavior, 43*, 29–50.

Clingingsmith, D., Khwaja, A. I., & Kremer, M. (2009). Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering. *Quarterly Journal of Economics, 124*(3), 1133–1170.

Cohen, J. (1989). Deliberation and Democratic Legitimacy. In A. Hamlin & P. Pettit (Eds.), *The Good Polity: Normative Analysis of the State* (pp. 17–34). Basil Blackwell.

Cook, S. W. (1978). "Interpersonal and Attitudinal Outcomes in Cooperating Interracial Groups. *Journal of Research and Development in Education, 12*(1), 97–113.

Dobson, A. (2012). Listening: The New Democratic Deficit. *Political Studies, 60*, 843–859.

Doherty, B. (2021). Couples Therapy Principles in the Design of the Braver Angels Red/Blue Workshop. Braver Angels. https://braverangels.org/couples-therapy-principles-in-the-design-of-the-braver-angels-red-blue-workshop/

Finnemore, M., & Sikkink, K. (1998). International norm dynamics and political change. *International Organization, 52*(4), 887–917.

Fishkin, James, Alice Siu, Larry Diamond and Norman Bradburn. 2021. "Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on "America in One Room"." *American Political Science Review* pp. 1–18.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: W. W. Norton & Company.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science, 14*, 265–287.

Glatter, Hayley. 2017. "The Most Polarized Freshmen Class in Half a Century." https://www.theatlantic.com/education/archive/2017/05/the-most-polarized-freshman-class-in-half-a-century/525135/

Gomila, R., & Clark, C. S. (2022). Missing data in experiments: Challenges and solutions. *Psychological Methods, 27*(2), 143–155.

Graham, M., & Svolik, M. (2020). Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States. *American Political Science Review, 114*(2), 392–409.

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential Validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006). *American Psychologist, 61*(1), 56–61.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Grossman, M., & Hopkins, D. (2015). Ideological Republicans and Group Interest Democrats: The Asymmetry of American Party Politics. *Perspectives on Politics, 13*(1), 119–139. https://doi.org/10.1093/poq/nfw001

Haidt, Jonathan and Marc J. Hetherington. 2012. "Look How Far We've Come Apart." *The New York Times* September 17.

Hartman, R., Blakey, W., Womick, J., Bail, C., Finkel, E. J., Han, H., Sar-, J., & rouf, Juliana Schroeder, Paschal Sheeran, Jay J. Van Bavel, Robb Willer and Kurt Gray. (2022). Interventions to Reduce Partisan Animosity. *Nature Human Behaviour, 6*(9), 1194–1205.

Hetherington, M. J., & Rudolph, T. J. (2015). *Why Washington Won't Work: Polarization, Political Trust, and the Governing Crisis*. University of Chicago Press.

Huddy, L., Mason, L., & Aarøe, L. (2015). Expressive Partisanship: Campaign Involve- ment, Political Emotion, and Partisan Identity. *American Political Science Review, 109*(1), 1–17.

Huges, A. (2017).https://www.pewresearch.org/short-reads/2017/05/17/5-facts-about-u-s-political-donations/

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly, 76*(3), 405–431.

Iyengar, S., & Krupenkin, M. (2018). The Strengthening of Partisan Affect. *Advances in Political Psychology, 39*(1), 201–218.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of affective Polarization in the United States. *Annual Review of Political Science, 22*, 129–146.

Jost, J. T. (2017). Presidential Address: Ideological Asymmetries and the Essence of Political Psychology. *Political Psychology, 38*(2), 167–208. https://doi.org/10.1111/pops.12407

Kalla, J. L., & Broockman, D. E. (2020). Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments. *American Political Science Review, 114*(2), 410–425.

Kalla, J. L., & Broockman, D. E. (2023). Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective- getting. *American Journal of Political Science, 67*(1), 185–204.

Kingzette, J., Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How Affective Polarization Undermines Support for Democratic Norms. *Public Opinion Quarterly, 85*(2), 663–677.

Kling, Jeffrey R., Jeffrey B. Liebman, Lawrence F. Katz and Lisa Sanbonmatsu. 2004. "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-Sufficiency and Health from a Randomized Housing Voucher Experiment." Social Science Research Network Working Paper No. 588942. http://papers.ssrn.com/abstract=588942

Levendusky, Matthew S. and Dominik A. Stecula. 2021. *We Need to Talk: How Cross Party Dialogue Reduces Affective Polarization*. Cambridge University Press.

Levitsky, S., & Ziblatt, D. (2018). *How Democracies Die*. Penguin Random House.

Mason, L. (2015). 'I Disrespectfully Agree: ' The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science, 59*(1), 128–145.

Mason, L. (2018). *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist, 62*(1), 16–42.

Mier, A. (2020). https://www.cnbc.com/2020/08/07/beyond-protests-college-students-donate-money-to-make-change-happen.html

Miller, N. (2002). Personalization and the Promise of Contact Theory. *Journal of Social Issues, 58*(2), 387–410.

Mousa, S. (2020). Building Social Cohesion between Christians and Muslims through Soccer in Post-ISIS Iraq. *Science, 369*(6505), 866–870.

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review, 113*(2), 517–529.

Mutz, Diana C. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press

Paler, L., Marshall, L., & Atallah, S. (2020). How Cross-Cutting Discussion Shapes Support for Ethnic Politics: Evidence from an Experiment in Lebanon. *Quarterly Journal of Political Science, 15*(1), 33–71. https://doi.org/10.1561/100.00018188

Paluck, E. L. (2010). Is It Better Not to Talk? Group Polarization, Extended Contact, and Perspective Taking in Eastern Democratic Republic of Congo. *Personality and Social Psychology Bulletin, 36*(9), 1170–1185.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice Reduction: Progress and Challenges. *Annual Review of Psychology, 72*(1), 533–560.

Paluck, E. L., Green, S. A., & Green, D. P. (2019). The Contact Hypothesis Re-evaluated. *Behavioural Public Policy, 3*(2), 129–158.

Pettigrew, T. F. (1998). Intergroup Contact Theory. *Annual Review of Psychology, 49*, 65–85.

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

Pew (2016). Pew Research Center's American Trends Panel, Wave 15 March & Wave 16 April, Combined Final Topline. https://www.people-press.org/wp-content/uploads/sites/4/2016/06/06-22-16-Partisanship-and-animosity-topline-for-release.pdf

Sanders, L. M. (1997). Against deliberation. *Political Theory, 25*(3), 347–376.

Santoro, Erik and David E Broockman. 2022. "The promise and pitfalls of cross-partisan conver- sations for reducing affective polarization: Evidence from randomized experiments." *Science advances* 8(25):eabn5515.

Santos, L. A., Voelkel, J. G., Willer, R., & Zaki, J. (2022). Belief in the utility of cross- partisan empathy reduces partisan animosity and facilitates political persuasion. *Psychological Science, 33*(9), 1557–1573.

Scacco, A., & Warren, S. S. (2018). Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria. *American Political Science Review, 112*(3), 654–677.

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science, 16*(2), 396–414.

Theuwis, Marie-Isabel, Carolien van Ham and Kristof Jacobs. 2021. "A Meta-Analysis of the Effects of Democratic Innovations on Citizens in Advanced Industrial Democracies." *ConstDelib Working Paper Series* 15:1–39.

Tuller, H. M., Bryan, C. J., Heyman, G. D., & Christenfeld, N. J. S. (2015). Seeing the other side: Perspective taking and the moderation of extremity. *Journal of Experimental Social Psychology, 59*, 18–23.

Voelkel, J. G., Michael, S., James, C., Sophia, L. P., Joseph, S. M., Chrystal, R., Isaias, G., Matthew, C., Dhaval, A., Levi, A. et al. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science, 386*(6719), eadh4764.

Webster, S. W., & Abramowitz, A. I. (2017). The Ideological Foundations of Affective Polarization in the U.S. Electorate. *American Politics Research, 45*(4), 621–647.

West, E. A., & Shanto, I. (2022). Partisanship as a Social Identity: Implications for Polarization. *Political Behavior, 44*(2), 807–838.

Wojcieszak, M., & Warner, B. R. (2020). Can interparty contact reduce affective polarization? A systematic test of different forms of intergroup contact. *Political Communication, 37*(6), 789–811.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Hannah Baron[1] · **Robert A. Blair**[2] · **Donghyun Danny Choi**[2] · **Laura Gamboa**[3] · **Jessica Gottlieb**[4] · **Amanda Lea Robinson**[5] · **Steven C. Rosenzweig**[6] · **Megan M. Turnbull**[7] · **Emily A. West**[8]

✉ Jessica Gottlieb
jagottlieb@uh.edu

Hannah Baron
hannahmbaron@gmail.com

Robert A. Blair
robert_blair@brown.edu

Donghyun Danny Choi
dannychoi@brown.edu

Laura Gamboa
lgamboa1@nd.edu

Amanda Lea Robinson
robinson.1012@osu.edu

Steven C. Rosenzweig
scrosen@bu.edu

Megan M. Turnbull
megan.turnbull@uga.edu

Emily A. West
eawest@pitt.edu

[1] Tulane University, New Orleans, USA

[2] Brown University, Providence, USA

[3] University of Notre Dame, Notre Dame, USA

[4] Hobby School of Public Affairs, University of Houston, Houston, USA

[5] Ohio State University, Columbus, USA

[6] Boston University, Boston, USA

[7] University of Georgia, Athens, USA

[8] University of Pittsburgh, Pittsburgh, USA