Can Americans Depolarize? Assessing the Effects of Reciprocal Group Reflection on Partisan Polarization

Hannah Baron¹ Robert A. Blair² Donghyun Danny Choi³
Laura Gamboa⁴ Jessica Gottlieb⁵ Amanda Lea Robinson⁶
Steven C. Rosenzweig⁷ Megan M. Turnbull⁸ Emily A. West⁹

Abstract

Overcoming America's deep partisan polarization poses a unique challenge: Americans must be able to disagree on policy 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-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 significantly 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.

¹Ph.D. Candidate in Political Science, Brown University

²Associate Professor of Political Science & International & Public Affairs, Brown University

³Assistant Professor of Political Science, University of Pittsburgh

⁴Assistant Professor of Political Science, University of Utah

⁵Associate Professor, Hobby School of Public Affairs, University of Houston

⁶Associate Professor of Political Science, Ohio State University

⁷Assistant Professor of Political Science, Boston University

⁸Assistant Professor of International Affairs, University of Georgia

⁹Assistant Professor of Political Science, University of Pittsburgh

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

1 Introduction

The American public is deeply polarized along partisan lines (Mason, 2018*b*; Iyengar et al., 2019). Americans express greater animosity toward those across the partisan divide than they did just a few decades ago (Haidt and Hetherington, 2012), and they increasingly view those on the other side of the aisle as untrustworthy, unpatriotic, and uninformed (Iyengar, Sood and Lelkes, 2012). “Affective polarization” of this sort can be problematic for democracy for at least three reasons. First, it can weaken electoral accountability by driving citizens to prioritize political ideology over democratic principles (Graham and Svobik, 2020; Kingzette et al., 2021), and to engage in motivated reasoning that desensitizes them to party performance (Freder, 2020; Little, Schnakenberg and Turner, forthcoming). Second, it can harm democratic legitimacy and the quality of democratic governance by encouraging legislative gridlock and disincentivizing compromise (Hetherington and Rudolph, 2015; McCoy, Rahman and Somer, 2018). Third, it can threaten democratic attitudes such as tolerance for the opposition (Levitsky and Ziblatt, 2019) and willingness to criticize co-partisans (Iyengar et al., 2019).

We report results from an experimental evaluation of an intervention that successfully reduced affective polarization by bringing balanced groups of Democrats and Republicans together for day-long workshops facilitated by trained, experienced moderators. The intervention aims to address one of democracy’s unique challenges: preserving the legitimacy of opposing beliefs while creating a foundation for civil exchange. The workshops are based on practices in couples counseling that seek to elicit mutual vulnerability and compassion while working to improve the health of a marriage

(Benson, McGinn and Christensen, 2012; Doherty, 2021). While Democrats and Republicans often observe each other during 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 in the workshops—not unlike the initial stages of couples counseling—focuses on generating understanding of “the other” and the motivations and experiences behind their beliefs.

The intervention highlights a new approach to overcoming affective polarization that is related to certain aspects of intergroup relations theories, such as contact 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 theories, it is distinct in that it does *not* aim for cooperation, consensus, or compromise. Instead, the goal is to promote the health of democratic norms and commitments by encouraging listening and understanding, which is believed to be important for the quality of deliberation (Dobson, 2012). The intervention seeks to promote participants’ ability to disagree with out-partisans while still humanizing them, understand their perspectives, and accept the legitimacy of their beliefs. These goals relate to work exploring ways to encourage listening, empathy, and perspective-taking/getting as a way to reduce intergroup hostility (Bruneau and Saxe, 2012).

Partnering with Braver Angels, a non-governmental organization, we evaluated the effects of this model 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, because they are quite literally the future of both the Republican and Democratic parties. Analysts have expressed particular concern about the heightened risk of polarization among youths who are coming of age at a time of partisan echo chambers in the home,

on campus, and on social media (Iyengar et al., 2019). That said, in Appendix B we show that our participants are comparable to the broader American electorate on many politically relevant characteristics, suggesting that our results may generalize beyond college campuses.

We find that the workshops significantly reduced both explicit and implicit measures of affective polarization. They 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—an especially striking finding given that the six months between the workshops and the endline were characterized by a pandemic, nationwide protests for racial justice, and a highly divisive presidential election. The effects of the workshops also appear to hold across partisan groups.¹

To better understand the mechanisms underlying the intervention's success, we systematically coded transcripts of the workshops. Through inductive analysis of these transcripts, we posit that the intervention was successful because it combined emotional with informational components, building empathy with the outgroup such that participants become more willing to update their priors about the outgroup in response to new information. Two unique features of the intervention facilitated these emotional and informational processes. First, the group-based nature of the perspective-getting exercises helped participants disassociate outgroup members from stereotypes, then re-associate new learnings about specific outgroup participants with the outgroup more generally. Second, the reciprocal nature of the exchange helped participants be mutually vulnerable with outgroup members. This emotional engagement appears to have facilitated information take-up. We call this model *reciprocal group reflection*, and argue that it is a potentially generalizable approach to depolarization.

¹We are under-powered in tests of heterogeneous treatment effects, however. We planned to host workshops at seven universities, but were able to conduct four due to the COVID-19 pandemic.

2 Affective Polarization: How Might Americans “Depolarize”?

Our outcome of interest—*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 (Abramowitz and Saunders, 2008; Iyengar et al., 2019; Webster and 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. The literature on partisan polarization in the US largely understands *affective polarization* as the result of the internalization of partisanship as a social identity, which generates strong ingroup preferences and outgroup bias (Huddy, Mason and Aarøe, 2015; West and Iyengar, 2020; Iyengar et al., 2019).

To be clear, we do not view *affective polarization* among citizens as the only, or even the greatest, threat to US democracy and other democracies today (Drutman, 2020; Levitsky and Ziblatt, 2019). We also do not believe that *affective polarization* among citizens is unconnected to elite polarization (Huddy and Yair, 2021). Indeed, one motivation for our study is that depolarization among citizens may encourage depolarization among elites as voters reward bipartisanship, compromise, and a shared commitment to democratic values. Although recent studies raise important questions about the causal effects of *affective polarization* on citizens’ commitment to democracy (Broockman, Kalla and Westwood, 2020; Costa, 2021), *affective polarization* may nevertheless undermine the quality of democratic governance over the long term in myriad ways, as discussed above.

2.1 Existing Attempts to Depolarize Citizens

Previous US-based depolarization interventions have sought to weaken partisan attachments, strengthen a common American identity, or shift perceptions of the outgroup, with mixed re-

sults. Interventions based in social psychology that employ self-affirmation or prime partisan ambivalence have not had the intended effect (Levendusky, 2018; West and Iyengar, 2020); interventions that signal warm relations between elites from different parties (Huddy and Yair, 2021) or correct misinformation about the composition of out-party members (Ahler and Sood, 2018) have been more promising. These findings suggest that correcting misperceptions of out-partisans (and possibly modeling friendly cross-partisan interactions) could be fruitful means of reducing affective polarization.

Other depolarization interventions are grounded in theories of intergroup contact (Allport, 1954) or deliberative democracy (Ryfe, 2005). While studies of intergroup contact find that it reduces discrimination in some spheres of social interaction, negative attitudes towards the outgroup tend to remain firmly entrenched (Mousa, 2020; Scacco and Warren, 2018). And while some studies find that deliberation may reduce affective polarization among partisans in the US (Fishkin et al., 2021; Levendusky and Stecula, 2021), others highlight the potentially perverse consequences of deliberation, including heightened out-group intolerance and antipathy (Mendelberg and Oleske, 2000; Paluck, 2010; Wojcieszak, 2011). Cross-partisan interactions on social media platforms have similarly been found to backfire (Bail et al., 2018).

2.2 A New Model: Reciprocal Group Reflection

Braver Angels “Red/Blue” workshops were designed to apply principles from evidence-based models of 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 consensus-building. Red/Blue workshops integrate components of previously successful depolarization efforts—e.g. provision of information about the policies and perspectives of out-partisans (Ahler and Sood, 2018)—with empathy-building exercises inspired by social psychology (Batson and Ahmad, 2009). The workshops thus combine cognitive and emotional mechanisms: they seek to reduce affective

Figure 1: Day-long Red/Blue Workshop

STRUCTURE OF THE RED/BLEU 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

polarization by providing information that counteracts narratives about out-partisans—e.g. that they are radical and dangerous (Berry and Sobieraj, 2014; Lelkes, Sood and Iyengar, 2017)—while simultaneously fostering out-party empathy and humanization (Martherus et al., 2019).

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. There are usually 5-8 workshop 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.

Workshop activities include a stereotypes exercise; a fishbowl exercise; lunch; a moderated Q&A; and a brainstorming session in which participants propose concrete steps to reduce affective polarization (Figure 1). 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.

In the “Stereotypes” exercise, Red and Blue participants break out into separate rooms with a moderator and are asked to brainstorm negative stereotypes about their own party. They then select five stereotypes and discuss why the stereotypes are largely false, but also attempt to identify a “kernel of truth” behind each. The groups reconvene to share their reflections; participants are asked to listen to out-partisans, but not to question or challenge what they say. In the “Fishbowl” exercise, each group takes turns sitting at the center of the room, surrounded by members of the outgroup. Participants in the inner circle are asked to share 1) why they think their party’s values and policies are good for the country, and 2) their uncertainties and reservations about their own party’s ideas. After both the Stereotypes and Fishbowl exercises, participants are asked to reflect on and discuss what they learned about the other side.

During the “Questions” exercise, participants again break out into their partisan groups and brainstorm questions they would like to ask the other side. Moderators help them craft questions that express genuine curiosity and a desire to understand the outgroup’s perspective; they guide participants away from “gotcha” questions that are intended to make a political point or catch outgroup members in a logical inconsistency. Each group is then instructed to ask their questions of the other side, and answer the other side’s questions. Finally, in the “How Can We Contribute” exercise, participants brainstorm actions they can take individually and collectively to reduce polarization in their communities, and share their ideas with the rest of the group.

These exercises are grounded in evidence-based approaches to couples therapy (Doherty, 2021). In particular, the Stereotypes and Fishbowl exercises aim to alter perceptions of the outgroup and guide participants away from one-sided explanations of the country’s problems. The Stereotypes,

Fishbowl, and Questions exercises elicit vulnerability as a way to increase empathy for and connection with the outgroup. The Questions exercise, ground rules, and moderator cues also seek to modify three common dysfunctional behaviors: refusing to listen, cherry-picking the positive aspects of the ingroup and negative aspects of the outgroup, and seeking opportunities to catch the other side making mistakes (e.g. through “gotcha” questions).

2.3 How Reciprocal Group Reflection May Reduce Affective Polarization

Through our observation of the workshops and coding and analysis of workshop transcripts, we identified two unique features of reciprocal group reflection that explain how it may reduce affective polarization and overcome the challenges that often undermine the potentially beneficial effects of inter-group contact (Pettigrew, 1998). The first, reciprocity, addresses the challenge that cognitive processes often must be married with emotional ones in order to alter perceptions; the second, group-based interaction, addresses the challenge that new inferences about outgroup members must extend beyond the individual if contact is to affect attitudes towards the outgroup as a whole.

To the first point, perspective giving and getting in our intervention is 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 an open mind. Importantly, this process does *not* involve deliberation, debate, or attempts at consensus-building, which may be counterproductive in some circumstances (Mendelberg and Oleske, 2000; Paluck, 2010; Fishkin et al., 2021).

A second unique characteristic of the Red/Blue model is that perspectives from the outgroup

are transmitted by a group rather than by a single individual. We argue that this helps generalize learnings from interpersonal contact—a process that involves what Pettigrew (1998) calls de-categorization, salient categorization, and re-categorization. Workshop activities foster empathy as participants listen to the perspectives of outgroup members, as in the Stereotypes exercise. In doing so, they engage in “de-categorizing” or disassociating that individual from the stereotype they hold of outgroup members. 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.

In addition to these more inductive observations and hypotheses, we pre-specified a set of mechanisms through which the Red/Blue workshop might reduce outgroup animus in our pre-analysis plan. These can also be grouped into cognitive and emotional channels. Cognitive mechanisms include 1) challenging negative stereotypes about the outgroup (Iyengar, Sood and Lelkes, 2012; Rothschild, Howat and Shafranek, 2019), 2) decreasing ideological extremity and thereby reducing the perception of threat from the outgroup (Abramowitz and Saunders, 2008; Rogowski and Sutherland, 2016; Webster and Abramowitz, 2017), 3) correcting misperceptions of mass ideological polarization or of the composition of the outgroup (Ahler and Sood, 2018; Mason, 2018*b*), and 4) countering the view that the ingroup has a monopoly on morality in politics (Garrett and Bankert, 2020).

Emotional mechanisms include 5) inducing empathy for the outgroup by witnessing their vulnerability and self-disclosure (Batson and Ahmad, 2009; Broockman and Kalla, 2016; Miller, 2002), 6) humanizing the outgroup by sharing higher-order emotions and opportunities for critical reflection (Cassese, 2021; Gubler, Halperin and Hirschberger, 2015; Martherus et al., 2019), and 7) weakening partisan attachments by reducing the salience of ingroup identity. These cognitive and emotional mechanisms are likely to reinforce one another. 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). Indeed, 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.

3 Research Design

3.1 Recruitment, Randomization, & Survey administration

We evaluated the impact of Red/Blue workshops during spring 2020 on four college campuses in the Northeast, Mid-Atlantic, South, and Midwest. Although our student sample is not representative of the US adult population, we address questions about generalizability by comparing our study participants to the broader American electorate in Appendix B, and by examining how our treatment effect estimates vary by theorized moderators (Krupnikov and Levine, 2014). To recruit 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. All undergraduate US citizens were eligible to apply.² We first solicited student expressions of interest via departmental and student group listservs. 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. Participants included a total of 165 subjects, including 116 “Blues” and 49 “Reds.”³ Of those who completed the baseline, 59 were randomly assigned to participate in

²We opted to focus on US citizens because we anticipated that non-citizens would be less likely to identify as either Democrats or Republicans, and less likely to exhibit affective polarization along these partisan lines.

³Braver Angels is a non-partisan organization with conservative and liberal leadership, and aims to recruit equal numbers of Reds and Blues for their Red/Blue workshops. However, recruitment on campuses more successfully

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. Members of the control group were invited to complete all follow-up surveys at the same time as the treatment group.

At two of the four universities, all Red-leaning students who completed the baseline were assigned to participate in the workshops, 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. At the other two universities, both Red- and Blue-leaning students were randomly assigned to treatment and control. Participant and observer roles within the treatment group were randomly assigned on the day of the workshop, stratifying by partisanship. (Because we had a small number of observers, we collapse these two categories for purposes of our analysis.) Per Braver Angels policy, each workshop had roughly the same number of participants from each party.

We report power calculations based on simulations using baseline values of our dependent variables in Appendix E. While our sample size is relatively 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. (Because we did not measure behavioral support for depolarization at baseline, we exclude this variable from our power analysis.) These would generally be considered small to moderate minimum detectable effects. Moreover, as we will see, our intention-to-treat (ITT)

attracted liberal-leaning than conservative-leaning students. This does not appear to be due only to the liberal skew on college campuses, as the tendency is apparent across the more conservative-leaning campuses in our sample. It is also consistent with the experiences of Braver Angels organizers nationally, who report greater difficulty recruiting Reds in both liberal- and conservative-leaning states. We speculate as to potential explanations for this discrepancy in the discussion.

estimates are remarkably consistent 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.

We carefully considered potential ethical implications before launching the project. To assess potential risks to participants, the principal investigators participated in workshops in their communities and engaged with Braver Angels national leadership on the design of the workshops. We also solicited information from previous workshop organizers on college campuses to better understand the particulars of running the intervention among a student population. We describe our risk assessment and the measures we took to minimize risks and maximize benefits in Appendix A. We recorded potentially negative participant experiences in the survey and found relatively few; among participants who reported feeling dissatisfied with the workshop, the most common reason was feeling sidelined by the moderator.

3.2 Measurement

Our primary outcome of interest is affective polarization, which we measure both explicitly and implicitly using our surveys. We measure an additional related outcome behaviorally: support for depolarization. We also use our surveys to construct indices corresponding to the pre-registered mechanisms described above (see Appendix C for their operationalization). 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;⁴ (2) a dummy for respondents who report identifying with the in-party primarily out of opposition to the out-party; (3) an index capturing the difference

⁴We 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. For example, if a Democrat reports a feeling of 80 towards other Democrats and a feeling of 30 towards Republicans, then she would receive a score of 50 on the thermometer.

between respondents' trust in the in-party and out-party;⁵ three-point Likert scales capturing how comfortable respondents would feel having outgroup members as (4) close personal friends and (5) neighbors, and (6) how comfortable they would feel if their best friend married an outgroup member; and (7) a dummy for respondents who believe the out-party represents a "serious threat" to the country. We create an index based on standardized responses to these seven questions, then re-standardize the midline and endline indices to their baseline values.

We measure *implicit* affective polarization using an implicit association test (IAT) administered in all three waves of the survey.⁶ 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, McGhee and Schwartz, 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.⁷ 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

⁵We 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.

⁶There 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.

⁷We increased this amount to \$20 at endline to mitigate attrition.

support for depolarization, i.e. respondents' willingness to take a 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.

Each of these three measures has strengths and weaknesses. The survey questions that comprise our index of explicit affective polarization are widely used in the field, but are potentially subject to experimenter demand effects. The IAT protocol used for our implicit measure of affective polarization is designed to overcome social desirability and other biases, but has been criticized on reliability and validity grounds (see, e.g., Brownstein, Madva and Gawronski, 2020)—though large-scale meta-analyses and replications suggest that implicit attitudes are pervasive, predictive of behavior, and highly correlated with explicit bias (Greenwald, Nosek and Sriram, 2006), and even critics characterize the IAT as most valid for measuring political attitudes (Schimmack, 2021). The behavioral measure leverages real, monetary outcomes to make misreporting more costly, but donation decisions may be a noisy proxy for political behavior. By employing three different outcome measures, we guard against the possibility that our results are driven by the particularities of any one. If the intervention affects the different measures in similar ways, this should increase our confidence that affective polarization has actually been changed.

3.3 Empirical Specification

Our pre-analysis plan specified a pooled analysis of midline and endline outcomes. We report these results in Appendix F, 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 baseline differences between the treatment and control groups. This specification is advantageous because it captures decay in the magnitude of our intention-to-treat (ITT) estimates over time, which is otherwise 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. 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). Standard errors are clustered by respondent.

4 Results

Figure 2 summarizes the ITT of the intervention on our three main outcomes at midline and endline, using the difference-and-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; treatment assignment was blocked on 1) campus, 2) party, and 3) affective polarization at baseline. A tabular version of the regression results used to compute marginal effects for Figure 2 is in Appendix I.

We observe a statistically significant negative ITT on each of our three main outcomes at midline; the effect is largest for our behavioral measure, smaller for our implicit measure, and smallest (and not quite statistically significant at the 5% level, $p = 0.063$) for the explicit measure. These effects decay over time for the explicit and implicit measures of affective polarization, losing statistical significance at endline, though the ITTs remain negative.⁸ The effect on our behavioral measure persists, remaining large and statistically significant even at endline. This is reassuring, as

⁸We present the pre-specified pooled results in Appendix F, where the ITT on the explicit measure is no longer statistically significant at conventional levels ($p = 0.141$) and the ITT on the implicit measure is weakly statistically significant ($p = 0.069$). The ITT on the behavioral measure remains strongly statistically significant regardless of specification.

Figure 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 are clustered by respondent.

the behavioral measure is arguably the one that is least susceptible to experimenter demand effects. The persistence of the behavioral ITT is especially striking given that the six months between the workshops and the endline witnessed the start of a pandemic, nationwide protests for racial equality, and a divisive presidential election.

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 and 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 Figure 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 1988 and 2016, the average out-party thermometer rating in the nationally-representative American National Election Studies (ANES) survey

Figure 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.

dropped from 45.54 to 25.99—a decline of 19.55 points (Iyengar and Krupenkin, 2018). In our midline, the average out-party thermometer rating among treatment group participants was 6.59 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-third of the decline in out-party “warmth” observed over nearly three decades. Similarly, between 1978 and 2008, the difference between in-party and out-party ratings in the ANES feeling thermometer grew by roughly 12.76 points (Broockman, Kalla and Westwood, 2020; Iyengar, Sood and Lelkes, 2012). In our midline, the average feeling thermometer difference among treatment group participants was 7.16 points smaller than the average difference among control group participants. If a treatment effect of this magnitude were extrapolated to the ANES sample, it would reverse more than half of the increase in affective polarization observed over three 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, Figure 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 Appendix G.

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 Appendix H. 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. The differences between attriters and non-attriters are otherwise generally small and not statistically significant, including predictors of attrition across these two groups.

Building on this analysis, in Appendix Figure F.3 we present attrition-adjusted treatment effect estimates using an inverse probability weighting (IPW) approach.⁹ Our treatment effect estimates using this procedure remain substantively similar in magnitude for all three of our main outcome measures, though the IPW-adjusted effects on explicit and implicit polarization at midline are significant only at the $p < 0.1$ 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 results. 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

⁹To generate IPWs, we regress attrition at midline and endline on the treatment indicator and interaction terms between the treatment indicator and baseline explicit polarization, implicit polarization, political engagement, and out-party contact.

Figure 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.

behaviors around polarization could easily apply to both groups. Yet despite similar measures at baseline, the groups diverge post-treatment, with the treated showing reduced levels of polarization relative to those in control.

Moreover, 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 and Peterson, 2019). The data suggests that our own study is in line with this overall trend.

5 Mechanisms

We find that Braver Angels Red/Blue workshops reduced affective polarization, at least in the short term. While our pre-analysis plan outlined seven potential mechanisms that might explain these results, as we show in Appendix C, 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 encouraging 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 relation to the seven pre-specified mechanisms. 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 Figure 5). 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 Figure 5, 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. Comments and conversations that recognize ingroup heterogeneity, show empathy for the outgroup, and 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 re-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 and group-based perspective-getting—two key features of the Braver Angels model—both appear to have played a role in facilitating these processes, as we discuss below.

5.1 Decategorization

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

Figure 5: Frequency of discussion topics

Note: Number of times each pre-specified mechanism appeared across all workshops. We exclude comments coded as individual ideological polarization. These reflect any ideological statement made by participants, and so are difficult to categorize as evidence of polarization or depolarization.

assumed. Comments signaling decreased ingroup identity salience were the third most common mechanism identified in the workshops (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. Separating 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 and 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 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 admit to 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 compelling 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 mutual empathy and understanding:

Ray (Blue): “I also found there to be, you know, lots of really good introspection on both sides. And I felt like really validated by a lot of the sort of 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.”

5.2 Salient Categorization

The disassociation mechanisms outlined above occurred alongside a process of salient categorization. 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 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. Not only did the intervention emphasize “Red” and “Blue” labels, framing ingroup and outgroup conversations in terms of “sides,” but it also fostered conversations among in-partisans that elicited group rather than individual traits.

5.3 Re-Categorization

Finally, we observe re-categorization when participants identify commonalities across groups. As depicted graphically in Figure 5, 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.”

6 Discussion

In an experimental evaluation of Braver Angels “Red/Blue” workshops designed to depolarize participants, we find statistically significant reductions in explicit and implicit measures of affective polarization, and a statistically significant increase in participants’ behavioral support for depolarization. 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 and Krupenkin, 2018). These effects appear to be consistent across Republican and Democratic participants.

Effects on the behavioral measure persist at endline; effects on the implicit and explicit measures remain negative, but lose statistical significance. The persistence of the behavioral effect is especially striking given that six months passed between the workshops and the endline—a period marked by the start of a pandemic, nationwide racial justice protests, and a hotly contested presidential election. 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 and Warren, 2018).

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).¹⁰ But participants also appear to have changed the way they *feel* about the outgroup, becoming more empathetic and more willing to humanize out-partisans.

Braver Angels' unique model appears to facilitate these cognitive and emotional processes

¹⁰This is often what we 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, Green and Green, 2019; Wojcieszak, 2011).

through group-level perspective-getting and reciprocity. Unlike many tests of the contact hypothesis, groups—rather than individuals—came into contact with each other through the workshops. Group-based reflection was helpful in translating lessons from individual outgroup members to generalized perceptions of the outgroup as a whole. As in evidence-based models of couples therapy, the expectation of reciprocity is important. Participants were willing to expose vulnerabilities knowing that outgroup members were expected to do the same. Moreover, unlike in mixed group deliberations, Red/Blue workshops are not focused on changing anyone’s mind, nor are they meant to correct misinformation. Previous studies show that people are more resistant to learning new information when they feel they are being manipulated or persuaded (Steindl et al., 2015), which Red/Blue workshop moderators explicitly discouraged.

Participants also enjoyed the support of ingroup members while being exposed to potentially challenging encounters with the outgroup. This may have increased participants’ openness to depolarization, in the same way that self-affirmation has been shown to reduce close-mindedness and inflexibility (Cohen et al., 2007). Taken together, these observations generate a new hypothesis for future testing: that new information (though not necessarily objective facts) about the outgroup is important for reducing affective depolarization, but that new information is most likely to have the intended effect if it is delivered in an environment that enables mutual humanization and empathy. 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 outgroup.

As for why the mechanisms identified in the qualitative data are not apparent in the survey data presented in Appendix C, we believe this discrepancy to be 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.

Are Red/Blue workshops a potentially viable solution to affective polarization nationwide? A full-day in-person workshop is admittedly hard to scale. However, Braver Angels offers a suite of shorter, virtual workshops that may have some of the same benefits. The discussion of mechanisms also highlights workshop features that can inform other, more scalable interventions. In particular, the model of group-based reciprocity is widely applicable. Many perspective-taking or -getting exercises are one-sided, whereas our study suggests these could offer more benefits if they were two-sided. In addition, perspective-getting often involves only one person's perspective, whereas our study suggests the importance of exchanging perspectives among groups.

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 (2019) 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. That said, there is accumulating evidence that non-threatening conversations can reduce intolerance (Kalla and 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.

References

- Abramowitz, Alan I. and Kyle L. Saunders. 2008. "Is Polarization a Myth?" *The Journal of Politics* 70(2):542–555.
- Ahler, Douglas J. and Gaurav Sood. 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. "Humanitarian Aid? Two Forms of Dehumanization and Willingness to Help after Natural Disasters." *British Journal of Social Psychology* 53:573–584.
- Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.
- Batson, Daniel C. and Nadia Y. Ahmad. 2009. "Using Empathy to Improve Intergroup Attitudes and Relations." *Social Issues and Policy Review* 3(1):141–177.
- Benson, Lisa A., Meghan M. McGinn and Andrew Christensen. 2012. "Common Principles of Couple Therapy." *Behavior Therapy* 43(1):25–35.
- Berry, Jeffrey M. and Sarah Sobieraj. 2014. *The Outrage Industry: Political Opinion Media and the New Incivility*. Oxford: Oxford University Press.
- Broockman, David E., Joshua L. Kalla and Sean J. Westwood. 2020. "Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not." OSF Preprint.
URL: [10.31219/osf.io/9btsq](https://doi.org/10.31219/osf.io/9btsq)
- Broockman, David and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352(6282):220–224.
- Brownstein, Michael, Alex Madva and Bertram Gawronski. 2020. "Understanding implicit bias: Putting the criticism into perspective." *Pacific Philosophical Quarterly* (101):276–307.
- Bruneau, Emile G. and Rebecca R. Saxe. 2012. "The Power of Being Heard: The Benefits of 'Perspective-giving' in the Context of Intergroup Conflict." *Journal of Experimental Social Psychology* 48(4):855–66.
- Cassese, Erin C. 2021. "Partisan dehumanization in American politics." *Political Behavior* 43:29–50.
- Cohen, Geoffrey L., David K. Sherman, Anthony Bastardi, Lillian Hsu, Michelle McGoe and Lee Ross. 2007. "Bridging the Partisan Divide: Self-Affirmation Reduces Ideological Closed-Mindedness and Inflexibility in Negotiation." *Journal of Personality and Social Psychology* 93(3):415–30.

- Cohen, Joshua. 1989. Deliberation and Democratic Legitimacy. In *The Good Polity: Normative Analysis of the State*, ed. Alan Hamlin and Philip Pettit. Oxford: Basil Blackwell pp. 17–34.
- Cook, Stuart W. 1978. “Interpersonal and Attitudinal Outcomes in Cooperating Interracial Groups.” *Journal of Research and Development in Education* 12(1):97–113.
- Costa, Mia. 2021. “Ideology, Not Affect: What Americans Want From Political Representation.” *American Journal of Political Science* 65:342–358.
- Dobson, Andrew. 2012. “Listening: The New Democratic Deficit.” *Political Studies* 60:843–859.
- Doherty, Bill. 2021. “Couples Therapy Principles in the Design of the Braver Angels Red/Blue Workshop.” Braver Angels.
URL: <https://braverangels.org/couples-therapy-principles-in-the-design-of-the-braver-angels-red-blue-workshop/>
- Druckman, James N. and Cindy D. Kam. 2011. *Student as Experimental Participants: A Defense of the "Narrow Data Base"*. New York: Cambridge University Press pp. 41–57.
- Drutman, Lee. 2020. *Breaking the Two Party Doom Loop: The Case for Multiparty Democracy in America*. New York: Oxford University Press.
- 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.
- Freder, Sean. 2020. “It’s No Longer the Economy, Stupid: Selective Perception and Attribution of Economic Outcomes.”
URL: https://seanfreeder.files.wordpress.com/2020/12/paper_economy.pdf
- Garrett, Kristin N. and Alexa Bankert. 2020. “The Moral Roots of Partisan Division: How Moral Conviction Heightens Affective Polarization.” *British Journal of Political Science* 50:621–640.
- Graham, Matthew and Milan Svolik. 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, Anthony G., Brian A. Nosek and N. Sriram. 2006. “Consequential Validity of the Implicit Association Test: Comment on Blanton and Jaccard (2006).” *American Psychologist* 61(1):56–61.
- Greenwald, Anthony G., Debbie E. McGhee and Jordan L.K. Schwartz. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.” *Journal of Personality and Social Psychology* 74(6):1464–1480.
- Gubler, Joshua Ronald, Eran Halperin and Gilad Hirschberger. 2015. “Humanizing the Outgroup in Contexts of Protracted Intergroup Conflict.” *Journal of Experimental Political Science* 1(1):36–46.

- Haidt, Jonathan and Marc J. Hetherington. 2012. "Look How Far We've Come Apart." *The New York Times* September 17.
URL: <https://campaignstops.blogs.nytimes.com/2012/09/17/look-how-far-weve-come-apart/>
- Hetherington, Marc J. and Thomas J. Rudolph. 2015. *Why Washington Won't Work: Polarization, Political Trust, and the Governing Crisis*. Chicago: University of Chicago Press.
- Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1):1–17.
- Huddy, Leonie and Omer Yair. 2021. "Reducing Affective Polarization: Warm Group Relations or Policy Compromise?" *Political Psychology* 42(2):291–309.
- Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76(3):405–431.
- Iyengar, Shanto and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Advances in Political Psychology* 39(1):201–218.
- Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J. Westwood. 2019. "The Origins and Consequences of affective Polarization in the United States." *Annual Review of Political Science* 22:129–146.
- Jost, John T., Jack Glaser, Arie W. Kruglanski and Frank J. Sulloway. 2003. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129(3):339–75.
- Kalla, Joshua L. and David E. Broockman. 2020. "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments." *American Political Science Review* 114(2):410–425.
- Kingzette, John, James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barr Ryan. 2021. "How Affective Polarization Undermines Support for Democratic Norms." *Public Opinion Quarterly* 85(2):663–677.
- Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(1):59–80.
- Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2017. "The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect." *American Journal of Political Science* 61:5–20.
- Levendusky, Matthew S. 2018. "When Efforts to Depolarize the Electorate Fail." *Public Opinion Quarterly* 82(3):583–592.
- Levendusky, Matthew S. and Dominik A. Stecula. 2021. *We Need to Talk: How Cross Party Dialogue Reduces Affective Polarization*. Cambridge University Press.
- Levitsky, Steven and Daniel Ziblatt. 2019. *How Democracies Die*. New York: Penguin Random House.

- Little, Andrew T., Keith E. Schnakenberg and Ian R. Turner. forthcoming. "Motivated Reasoning and Democratic Accountability." *American Political Science Review* .
URL: <https://osf.io/preprints/socarxiv/esfy6/>
- Martherus, James L., Andres G. Martinez, Paul K. Piff and Alexander G. Theodoridis. 2019. "Party Animals? Extreme Partisan Polarization and Dehumanization." *Political Behavior* Forthcoming.
- Mason, Lilliana. 2018a. "Losing Common Ground: Social Sorting and Polarization." *The Forum* 16(1):47–66.
- Mason, Lilliana. 2018b. *Uncivil Agreement: How Politics Became Our Identity*. Chicago: University of Chicago Press.
- McCoy, Jennifer, Tahmina Rahman and Murat Somer. 2018. "Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities." *American Behavioral Scientist* 62(1):16–42.
- Mendelberg, Tali and John Oleske. 2000. "Race and Public Deliberation." *Political Communication* 17(2):169–191.
- Miller, Norman. 2002. "Personalization and the Promise of Contact Theory." *Journal of Social Issues* 58(2):387–410.
- Mousa, Salma. 2020. "Building Social Cohesion between Christians and Muslims through Soccer in Post-ISIS Iraq." *Science* 369(6505):866–870.
- Mummolo, Jonathan and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113(2):517–529.
- Paluck, Elizabeth Levy. 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, Elizabeth Levy, Roni Porat, Chelsey S. Clark and Donald P. Green. 2021. "Prejudice Reduction: Progress and Challenges." *Annual Review of Psychology* 72(1):533–560.
- Paluck, Elizabeth Levy, Seth A. Green and Donald P. Green. 2019. "The Contact Hypothesis Re-evaluated." *Behavioural Public Policy* 3(2):129–158.
- Pettigrew, Thomas F. 1998. "Intergroup Contact Theory." *Annual Review of Psychology* 49:65–85.
- Pew. 2014. "Political Polarization in the American Public.".
URL: <https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/>
- Rogowski, Jon C. and Joseph Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38(2):485–508.

- Rothschild, Jacob E., Adam J. Howat and Richard M. Shafranek. 2019. "Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization." *Political Behavior* 41(2):423–443.
- Ryfe, David M. 2005. "Does Deliberative Democracy Work?" *Annual Review of Political Science* 8(1):49–71.
- Scacco, Alexandra and Shana S. Warren. 2018. "Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria." *American Political Science Review* 112(3):654–677.
- Schimmack, Ulrich. 2021. "The Implicit Association Test: A Method in Search of a Construct." *Perspectives on Psychological Science* 16(2):396–414.
- Simas, Elizabeth N., Scott Clifford and Justin H. Kirkland. 2020. "How empathic concern fuels political polarization." *American Political Science Review* 114(1):285–269.
URL: <https://doi.org/10.1017/S0003055419000534>
- Steindl, Christina, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch and Jeff Greenberg. 2015. "Understanding Psychological Reactance." *Zeitschrift Fur Psychologie* 223(4):205–214.
- Webster, Steven W. and Alan I. Abramowitz. 2017. "The Ideological Foundations of Affective Polarization in the U.S. Electorate." *American Politics Research* 45(4):621–647.
- West, Emily A. and Shanto Iyengar. 2020. "Partisanship as a Social Identity: Implications for Polarization." *Political Behavior* Forthcoming.
- Wojcieszak, Magdalena. 2011. "Deliberation and Attitude Polarization." *Journal of Communication* 61(4):596–617.

Supporting Information for *Can Americans Depolarize?*

Table of Contents

A Ethical Considerations	1
B Generalizability	2
C Pre-Specified Mechanisms	4
D Descriptive Statistics	5
E Power Calculations	7
F Main Results Pooled across Rounds	7
G Heterogeneous Treatment Effects	8
H Attrition	8
I Supplemental Analysis Tables	8

A Ethical Considerations

In this section, we first describe our risk assessment process, and the risks and benefits we identified of doing research on affective polarization on college campuses. Second, we discuss our methods to minimize potential risk and maximize potential benefits. Third, we discuss the informed consent process and participant remuneration. Fourth, we evaluate our ethical considerations based on participant feedback and coding of workshop transcripts. Last, we describe the instruments and data that are publicly available upon publication. We append a copy of our informed consent to this section.

We developed our risk assessment based on consultations with Braver Angels leadership and our own participation in Red/Blue workshops in our respective communities. Based on these exchanges, we identified emotional distress and the violation of privacy as the two potential risks of participating in the workshop-based study. We anticipated potential sources of emotional distress such as 1) discomfort while discussing opposing political attitudes and, at an unlikely extreme, 2) exposure to hateful speech or discrimination while discussing political issues. We also anticipated that even though the informed consent process and ground rules before each workshop ask participants to respect others' privacy, participants may transgress these ground rules and share potentially sensitive and identifying information about one another after the intervention.

While there were no direct, personal benefits for participation in the Braver Angels workshops, there were a number of potential indirect benefits. These include feeling more understood and respected by others, and feeling empowered to communicate more effectively with out-partisans who happen to hold opposing views. To minimize potential harms and maximize potential benefits, we coordinated with Braver Angels leadership to identify experienced moderators to lead the campus workshops. All moderators are trained to explain and enforce ground rules of mutual tolerance and respect. With these measures in place, we judged that the potential benefits to participation in the workshops would outweigh the potential risks. We also made potential risks clear in the informed consent process.

In the initial call for expressions of interest via university listservs, potential participants were provided descriptions of and links to Braver Angels programming, and were informed that participating in the workshop meant participating in an academic study. Before completing the baseline survey, which was self-administered on Qualtrics, participants were presented with a more complete informed consent script reiterating that participation in the workshop implied participation in an academic study involving multiple rounds of surveys, and that participation in the workshop would be randomly assigned. The informed consent is copied below. Respondents received a \$5 Amazon gift card for the baseline survey and a \$10 Amazon gift card for the midline survey; we increased this amount to \$20 at endline to mitigate attrition.

We used two measurement strategies to evaluate the efficacy of the informed consent process and gauge potential unintended consequences of the workshops. First, at the end of the midline survey, we asked participants if they were satisfied or dissatisfied with their participation in the workshops (and the study), and why. The latter was an open-ended question, allowing respondents to express reasons for satisfaction or dissatisfaction that we might not have been able to anticipate. Second, we coded the qualitative data for evidence of unintended consequences of workshop participation.

Out of the 121 respondents who completed the midline, only five expressed dissatisfaction with the workshops. Two of them reported feeling disappointed that they had been assigned the relatively passive role of observers, instead of being assigned to the active participant role. The other three (two Reds and one Blue) expressed dissatisfaction either because their values were not represented in the survey, or because they perceived that the workshop structure did not allow the group to work through important policy issues. In coded transcripts from all four workshops, we found a total of 20 instances of unintended consequences. Specifically, there were four instances where two participants total expressed a reluctance to participate (one Red and one Blue), five occurrences of four participants total expressing feelings of stigmatization within the workshop (three Reds and one unidentified participant), and 11 occasions where 11 participants were skipped over or dismissed by the moderator (six Reds and five Blues).

B Generalizability

Although our student sample does not constitute a representative sample of the American public, study participants are comparable to the broader electorate on many politically-relevant characteristics and, importantly, our sample varies on theorized moderators (Krupnikov and Levine, 2014). According to Druckman and Kam, convenience samples are less problematic with regards to ex-

ternal validity when there is variance along theorized moderators that allow researchers to explore how treatment effects vary along attributes considered to be relevant in the general population Druckman and Kam (2011),. To recap, our pre-analysis plan identified party ID (Jost et al., 2003), partisan extremity (Levendusky, 2018; Rogowski and Sutherland, 2016; Webster and Abramowitz, 2017), political engagement (Pew, 2014), prior contact with the outgroup (Mason, 2018a), and dispositional empathy (Simas, Clifford and Kirkland, 2020) as potential moderators.

First, although our sample varies by party ID, it skews more heavily Democratic compared to the nationally-representative American National Election Studies (ANES) survey. Roughly 56% of our sample self-identifies as Democrats, 15% as Independents, and 24% as Republicans, compared to 42%, 17%, and 37%, respectively, among ANES 2019 respondents (and compared to 46%, 14%, and 39%, respectively, among ANES 2016 respondents).

Second, when it comes to partisan extremity, Democrats and Republicans in our sample are only marginally more ideologically extreme than their ANES 2016 counterparts. Mean ideology among Democrats in the ANES is 3 (on their 7 point scale), while mean ideology among Democrats in our sample is 2.14. The difference is even smaller among Republicans. In the ANES Republicans have a mean ideology score of 5.36; in our sample they have a mean ideology score of 5.78. Our sample also varies by partisan extremity. As per our pre-analysis plan, our measure of partisan extremity is the the absolute value of the difference between respondents' self-placement on a 7-point (1-7) political ideology scale and the middle/moderate value (4). Roughly 8% identify as moderate/middle of the road, 24% slightly liberal or conservative, 52% as either liberal or conservative, and 17% as either extremely liberal or extremely conservative.

Third, although our measures for political engagement differ from those in the ANES, in general, our respondents appear to be only somewhat more politically engaged than the average ANES 2019 respondent. Roughly 49% of our sample reports "following news about the government" very closely, 45% somewhat closely, and 6% not very closely; approximately 59% discuss politics with someone every day, 35% a few times a week, and 7% a few times a month. By comparison, roughly 56% of ANES 2019 respondents report "following what's going on" most of the time, 27% some of the time, 9% every now and then, and 7% hardly at all. (Political engagement is lower, on average, in ANES 2016.) Perhaps unsurprisingly given the perceived stakes on both sides surrounding the 2020 presidential election, about 97% of our sample expressed their intention to vote in the upcoming election. While the ANES does not ask about respondents' intention to vote in upcoming elections, of those who were asked in 2019 if they had voted in the 2016 presidential election, roughly 81% reported that they had while 16% said that they had not and about 3% were unsure.

Our other questions about political engagement and activity do not greatly overlap with those in the 2019 ANES, but importantly, our sample varies across these measures. Over the past year, approximately 55% reported attending a rally (7% in the 2019 ANES),¹¹ compared to 45% (93% in the ANES) who did not; roughly 32% had volunteered for a political campaign, party, or candidate, whereas 68% had not; about 34% of our sample (20% in the ANES) reported donating money to a political campaign or party whereas about 67% of our sample (80% in the ANES) did not; roughly 37% of our sample were active members of a political organization and 67% were not; 45% of sample reported contacting an elected official compared to 55% that had not; approximately

¹¹ All the following ANES figures are from 2019 unless otherwise stated.

56% of our sample (13% in the ANES) had displayed some kind of a political message whereas 44% of our sample (87% in the ANES) had not; 55% of our sample (34% in the ANES) expressed support for a political campaign on social media while 45% of our sample (67% in the ANES) did not; roughly 44% of our sample reported attending a protest while about 56% did not; and finally, approximately 64% of our sample said they signed a petition whereas 36% did not. In short, our sample varies significantly when it comes to different modes of political engagement which is an important consideration regarding its external validity (Druckman and Kam, 2011).

Although our last two moderators - prior outgroup contact and dispositional empathy - do not appear in the ANES (rendering us unable to say how our sample compares to American adults), they again vary across our sample. Roughly 1% said they interacted with members of the outparty less than once a year, and another 1% reported about once a year. Approximately 2% interacted with the outgroup once every six months, 7% on a monthly basis, 32% on a weekly basis, and 56% on a daily basis. When asked how often they talked politics with the outparty, approximately 4% reported less than once a year, 3% once a year, 10% once every six months, 30% monthly, 41% weekly, and 12% daily. Our survey also asked respondents about the partisanship of their close friends. About 50% reported that nearly all of their close friends were co-partisans, 40% said that half of their close friends were co-partisans, 10% just a few, and .6% reported none. About 16% said that they had zero close friends from the outparty, 59% just a few, 22% about half, and 3% almost all. When asked about close friends who were independents, 14% reported none, 65% just a few, 16% about half, and 5% almost all.

We also see variation in general empathy in our sample, our last moderator. Our pre-analysis plan specified an index based on five questions from a larger battery to capture dispositional empathy (Simas, Clifford and Kirkland, 2020). For each question, respondents were asked how well the following statement described them, and were given a scale from not very well (-2) to very well (2). For the empathy index, the sample mean is -.008 with a standard deviation of .65, and a minimum value of -2.1 and a maximum of 1.12.

Finally, and perhaps most importantly, while our sample is less affectively polarized than the ANES 2016 sample, the differences are relatively small.¹² Our respondents express more favorable views of both the in-party and the out-party than the ANES sample, with an average in-party thermometer rating of 72 in our sample (compared to 67 in ANES), and an average out-party thermometer rating of 37 (compared to 26 in ANES). But because *both* the in-party and out-party thermometer ratings are higher in our sample, the difference between them—one of our key measures of explicit affective polarization—is not much lower than in ANES. The average difference between the in-party and out-party thermometer ratings in our sample is 35, compared to 41 in ANES. This suggests that our results are unlikely to be artifacts of atypically low levels of pre-treatment affective polarization in our sample.

C Pre-Specified Mechanisms

As noted in the main body of the paper, we find little evidence in the quantitative data for the pre-specified measures for seven different mechanisms. We use the survey data to construct indices corresponding to the mechanisms described in the paper and PAP: (1) outgroup stereotyping,

¹²Data on affective polarization is not available in ANES 2019.

measured as the difference between respondents' positive and negative stereotypes of the in-party and out-party (e.g. the extent to which they view the in-party as more patriotic than the out-party, or the out-party as more selfish than the in-party); (2) perceived mass ideological polarization, capturing the extent to which respondents believe the public is polarized around these same issues; (3) outgroup empathy, capturing the extent to which respondents report being willing and able to empathize with the out-party's perspective; (4) outgroup humanization, capturing the extent to which respondents use (de)humanizing language to characterize the outgroup (e.g. by describing outgroup members as incapable of feeling love and compassion); (5) moralization of politics, capturing the extent to which respondents frame their views on contentious political issues in terms of right and wrong; (6) ingroup identity salience, capturing the extent to which respondents characterize their membership in the ingroup as an important part of their identity; and (7) ideological polarization, capturing the extent to which respondents take extreme positions on five political and social issues weighted by the importance they place on those issues¹³. All of these indices are standardized to their baseline values. Figure S.1 below summarizes the ITT on each of these seven mechanisms.

As shown in Figure S.1, the only statistically significant effect of the treatment on any of the mechanism measures is a reduction of the moralization of politics in the endline survey.

D Descriptive Statistics

Tables C.1 through C.3 report descriptive statistics.

Table C.1: Outcome measures

Variable	N	Mean	SD	Min	Max
Explicit	406	0.00	1.02	-2.43	2.38
Implicit	391	-0.05	1.00	-3.46	2.20
Behavioral	238	-0.44	1.46	-5.38	0.31

Table C.2: Student demographics

	Gender		
	Control	Treatment	Total
Female	20 (51%)	26 (42%)	46 (46%)
Male	19 (49%)	34 (55%)	53 (52%)
Other/NA	0 (0%)	2 (3%)	2 (2%)
Total	39	62	101

	Ethnicity		
	Control	Treatment	Total
Non-White	12 (31%)	11 (18%)	23 (23%)
White	27 (69%)	51 (82%)	78 (77%)
Total	39	62	101

¹³We operationalize ideological polarization using the issue extremity measure from Mason (2018b)

Figure S.1: Treatment effects on potential mechanisms

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Table C.3: Workshop participant demographics
(assigned to treatment only)

Gender			
	Non-Participant	Participants	Total
Female	1 (100%)	22 (43%)	23 (44%)
Male	0 (0%)	27 (53%)	27 (52%)
Other/NA	0 (0%)	2 (4%)	2 (4%)
Total	1	51	52

Ethnicity			
	Non-Participant	Participants	Total
Non-White	0 (0%)	9 (18%)	9 (17%)
White	1 (100%)	42 (82%)	43 (83%)
Total	1	51	52

E Power Calculations

We simulate the power of our design using our baseline measures of explicit and implicit affective polarization. (Since we did not measure behavioral support for depolarization at baseline, we exclude this variable from our power analysis.) We first simulate our randomization procedure as it was implemented 1,000 times. In each round of the simulation, we regress baseline affective polarization on an indicator for treatment assignment with block fixed effects. The simulation generates a distribution of 1,000 intention-to-treat (ITT) estimates under the null hypothesis of no effect (since in expectation there should be no difference in affective polarization between the treatment and control groups at baseline). We then calculate the standard deviation of this distribution (which approximates the standard error of our estimator), and multiply that standard deviation by 2.8 (the rule of thumb proposed by Gelman and Hill 2006 for 80% power). Based on this procedure, our minimum detectable effect (MDE) on explicit affective polarization is approximately 0.21 standard deviations, and our MDE on implicit affective polarization is approximately 0.46 standard deviations. These would generally be considered small to moderate effect sizes.

F Main Results Pooled across Rounds

Figure D.1 reports the effects of the workshops on affective polarization and support for depolarization, pooling midline and endline responses.

G Heterogeneous Treatment Effects

Figures E.1 through E.4 report heterogeneous treatment effects by baseline ideological polarization, baseline political engagement, baseline rate of out-party contact, and baseline dispositional empathy.

H Attrition

It is possible that attrition between the baseline survey and the midline and endline surveys may bias our treatment effect estimates. To investigate this possibility, we first examine attrition in midline and endline responses as a function of observed pre-treatment covariates reported in the baseline survey. The single most important predictor of attrition is treatment assignment. Respondents who were assigned to treatment were 22 percentage points and 11 percentage points more likely to participate in the midline and endline surveys, respectively. For the midline, the only other pre-treatment covariate that predicts attrition at the 10% level is perceived mass ideological polarization. For the endline, two pre-treatment covariates—implicit affective polarization and ideological constraint—are significant predictors of attrition. Otherwise attriters and non-attriters are statistically indistinguishable on observables.

Despite these apparently minor differences in the full sample, it is also possible that pre-treatment covariates differentially affect attrition in the treatment and control conditions. To probe this possibility, in Figures F.1 and F.2 we report the marginal effects of these covariates on attrition, disaggregated by treatment assignment. At midline, there appears to be some heterogeneity in the relationship between baseline covariates and attrition by treatment status. Specifically, pre-treatment levels of political engagement and out-party contact appear to differentially predict attrition in the treatment and control groups. But these are exceptions; in general the differences between groups are small and not statistically significant at conventional levels. At endline, political engagement, partisan feeling thermometer gap, and negative partisanship appear to differentially predict attrition, but again, these are exceptions.

Finally, in Figure F.3 we present attrition-adjusted treatment effect estimates on our three main outcome variables using an inverse probability weighting (IPW) approach. To generate IPWs, we regress dummies for attrition at midline and endline on treatment status and the interaction between treatment status and baseline explicit polarization, implicit polarization, political engagement, and out-party contact. Our treatment effect estimates remain substantively similar in magnitude for all three outcome measures, though the effects on explicit and implicit polarization at midline are significant only at the $p < 0.1$ level after applying IPW.

I Supplemental Analysis Tables

Tables G.1 through G.5 replicate the figures in the paper in tabular form.

Figure D.1: Treatment effects on affective polarization and support for depolarization, pooled across rounds

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure E.1: Treatment effects on affective polarization and support for depolarization by baseline ideological polarization

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure E.2: Treatment effects on affective polarization and support for depolarization by baseline political engagement

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure E.3: Treatment effects on affective polarization and support for depolarization by baseline rate of out-party contact

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure E.4: Treatment effects on affective polarization and support for depolarization by baseline dispositional empathy

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure F.1: Pretreatment predictors of midline attrition by treatment status

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Figure F.2: Pretreatment predictors of endline attrition by treatment status

Note: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Table F.1: Predictors of Attrition: Midline and Endline

	(1)		(2)	
	Midline Attrition		Endline Attrition	
Treatment	-0.338	(-1.27)	-0.0985	(-0.44)
Explicit Affective polarization	0.214	(0.53)	0.347	(1.06)
Implicit Affective Polarization	-0.0467	(-0.54)	-0.0784	(-1.14)
Partisan Feeling Thermometer Gap	-0.197	(-1.38)	-0.207	(-1.50)
Negative Partisanship	-0.160	(-1.47)	-0.0714	(-0.55)
Partisan Trust Gap	-0.0282	(-0.17)	-0.0779	(-0.55)
Comfort w/ Out Party Friends	0.118	(0.76)	-0.0848	(-0.56)
Comfort w/ Out Party Neighbors	-0.215	(-1.73)	-0.142	(-1.08)
Comfort w/ Out Party Marriage	-0.0373	(-0.33)	-0.153	(-1.50)
Outgroup Stereotyping	0.0128	(0.13)	0.0735	(0.71)
Ideological Constraint	-0.111	(-1.10)	-0.206	(-1.98)
Ideological Extremity	0.0425	(0.40)	-0.128	(-1.11)
Perceived Mass Ideo. Polarization	0.0579	(0.75)	-0.0528	(-0.57)
Outgroup Empathy	0.120	(1.19)	-0.137	(-1.48)
Outgroup Humanization	0.0589	(0.79)	-0.0225	(-0.24)
Moralization of Politics	0.143*	(2.06)	0.0961	(1.01)
Ingroup Identity Salience	0.00537	(0.05)	0.0442	(0.51)
Extremity of Ideological Self-Placement	-0.163	(-1.66)	0.00945	(0.12)
Political Engagement	-0.0852	(-0.45)	0.183	(1.06)
Outparty contact	-0.0458	(-0.31)	-0.225	(-1.28)
Dispositional empathy	-0.219	(-1.33)	0.271	(1.75)
Treatment $\times$ Explicit Affective polarization	-0.451	(-0.94)	-0.836	(-1.72)
Treatment $\times$ Implicit Affective Polarization	0.0217	(0.19)	-0.0682	(-0.81)
Treatment $\times$ Partisan Feeling Thermometer Gap	0.316	(1.62)	0.516*	(2.62)
Treatment $\times$ Negative Partisanship	0.314*	(2.22)	0.238	(1.39)
Treatment $\times$ Partisan Trust Gap	0.0221	(0.11)	0.0526	(0.30)
Treatment $\times$ Comfort w/ Out Party Friends	0.0286	(0.15)	0.165	(0.77)
Treatment $\times$ Comfort w/ Out Party Neighbors	0.362	(1.93)	0.369*	(2.11)
Treatment $\times$ Comfort w/ Out Party Marriage	-0.0516	(-0.35)	0.203	(1.43)
Treatment $\times$ Outgroup Stereotyping	-0.101	(-0.77)	-0.251*	(-2.12)
Treatment $\times$ Ideological Constraint	0.0545	(0.44)	0.255	(1.90)
Treatment $\times$ Ideological Extremity	-0.0376	(-0.30)	0.0125	(0.09)
Treatment $\times$ Perceived Mass Ideo. Polarization	0.0347	(0.27)	0.265	(1.97)
Treatment $\times$ Outgroup Empathy	-0.116	(-1.00)	0.233*	(2.16)
Treatment $\times$ Outgroup Humanization	-0.132	(-1.62)	0.0952	(0.92)
Treatment $\times$ Moralization of Politics	-0.0636	(-0.71)	-0.101	(-0.93)
Treatment $\times$ Ingroup Identity Salience	0.0955	(0.86)	-0.0166	(-0.17)
Treatment $\times$ Extremity of Ideological Self-Placement	0.0603	(0.42)	-0.00552	(-0.05)
Treatment $\times$ Political Engagement	0.0971	(0.37)	0.0194	(0.09)
Treatment $\times$ Outparty contact	0.0444	(0.28)	0.202	(1.07)
Treatment $\times$ Dispositional empathy	0.183	(0.96)	-0.432*	(-2.09)
Constant	0.985*	(2.39)	0.0668	(0.28)
Observations	139		139	
Adjusted R^2	0.169		0.121	
F Statistic	5.196		2.977	

Notes: t statistics in parentheses. All variables used are from the baseline survey. Specifications include block fixed effects. Coefficients for the block fixed effects are omitted. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure F.3: Treatment effects on affective polarization and support for depolarization using IPW for attrition

Note: Models to generate predicted probabilities for the IPW analysis included treatment status, and explicit polarization, implicit polarization (outcomes measured at baseline) and index of baseline political engagement and out-party contact, which showed some suggested evidence of differential attrition across treatment and control, interacted with the treatment indicator. Adjusted treatment effect estimate are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization.

Table G.1: Treatment effects on affective polarization and support for depolarization

	(1) Explicit	(2) Implicit	(3) Behavioral
Treatment	-0.009 [0.116]	-0.039 [0.165]	-0.826*** [0.283]
Midline	0.142 [0.092]	0.170 [0.168]	
Endline	0.021 [0.134]	-0.084 [0.179]	-0.073 [0.165]
Treatment $\times$ midline	-0.289** [0.133]	-0.394* [0.201]	
Treatment $\times$ endline	-0.105 [0.172]	-0.258 [0.219]	0.047 [0.311]
Observations	351	341	208

Notes: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization. Standard errors are clustered by respondent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G.2: Marginal effects on affective polarization and support for depolarization

	(1) Explicit	(2) Implicit	(3) Behavioral
Midline	-0.298* [0.159]	-0.433** [0.176]	-0.826*** [0.283]
Endline	-0.114 [0.184]	-0.297 [0.203]	-0.779*** [0.271]
Observations	351	341	208

Notes: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization. Standard errors are clustered by respondent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G.3: Treatment effects on affective polarization and support for depolarization, in-party vs. out-party

	(1) In-party trust	(2) Out-party trust	(3) In-party thermometer	(4) Out-party thermometer
Treatment	-0.027 [0.158]	0.173 [0.140]	0.016 [0.171]	-0.009 [0.147]
Midline	0.137 [0.133]	-0.300** [0.117]	-0.185 [0.167]	-0.156 [0.129]
Endline	0.141 [0.183]	-0.271* [0.151]	-0.375** [0.172]	-0.189 [0.132]
Treatment $\times$ midline	-0.010 [0.167]	0.177 [0.182]	0.214 [0.192]	0.367** [0.174]
Treatment $\times$ endline	-0.234 [0.240]	-0.045 [0.198]	0.310 [0.214]	0.070 [0.179]
Observations	351	351	351	351

Notes: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization. Standard errors are clustered by respondent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G.4: Marginal effects on affective polarization and support for depolarization, in-party vs. out-party

	(1) In-party trust	(2) Out-party trust	(3) In-party thermometer	(4) Out-party thermometer
Midline	-0.037 [0.206]	0.350** [0.169]	0.230 [0.220]	0.358* [0.182]
Endline	-0.260 [0.238]	0.127 [0.202]	0.326 [0.233]	0.061 [0.173]
Observations	351	351	351	351

Notes: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization. Standard errors are clustered by respondent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table G.5: Treatment effects on affective polarization and support for depolarization by party

	(1) Explicit	(2) Implicit	(3) Behavioral
Treatment	-0.057 [0.161]	0.038 [0.199]	-0.777** [0.377]
Midline	0.119 [0.129]	0.284 [0.197]	
Endline	0.249* [0.145]	0.038 [0.254]	0.083 [0.280]
Treatment $\times$ midline	-0.154 [0.187]	-0.491* [0.257]	
Treatment $\times$ endline	-0.137 [0.198]	-0.363 [0.309]	-0.144 [0.437]
Republicans	-0.209 [0.319]	1.576*** [0.388]	-1.483** [0.620]
Independents	-0.076 [0.236]	-0.134 [0.298]	0.317 [0.338]
Treatment $\times$ Republicans	0.171 [0.246]	-0.834** [0.390]	0.452 [0.778]
Treatment $\times$ Independents	-0.031 [0.308]	0.259 [0.409]	-0.638 [0.710]
Midline $\times$ Republicans	-0.087 [0.184]	-0.730 [0.588]	
Midline $\times$ Independents	0.198 [0.303]	-0.001 [0.375]	
Endline $\times$ Republicans	-0.655** [0.283]	-0.929** [0.387]	-0.479 [0.359]
Endline $\times$ Independents	-0.374 [0.361]	0.136 [0.457]	-0.433 [0.324]
Treatment $\times$ midline $\times$ Republicans	-0.181 [0.295]	0.834 [0.657]	
Treatment $\times$ midline $\times$ Independents	-0.372 [0.406]	-0.139 [0.496]	
Treatment $\times$ endline $\times$ Republicans	0.074 [0.344]	1.100** [0.515]	0.317 [0.884]
Treatment $\times$ endline $\times$ Independents	0.114 [0.515]	-0.276 [0.654]	0.802 [0.853]
Observations	351	341	208

Notes: Estimates are standardized and include block fixed effects. Treatment assignment was blocked on 1) campus, 2) party, and 3) baseline affective polarization. Standard errors are clustered by respondent. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.