Reducing Meta-Perceptions as a Field Test of Depolarization Initiatives[*]

Daniel B. Markovits[†] Aaron Christensen[‡] Andrew I. Thompson[§]

August 31, 2025

Abstract

Correcting false and negative beliefs about political opponents has shown promise in reducing anti-democratic attitudes and polarization. Despite the simple nature of such corrections, there is little extant evidence that they are effective beyond immediately administered survey outcomes and it is unclear which voters opt-in to such interventions. To test these mechanisms, we worked with a partner organizations to implement a depolarization initiative that bundled factual belief corrections with elites modeling civil disagreement. We recruited an online panel of 3,461 eligible respondents and then randomized an offer to attend a 30 minute depolarization event in which bipartisan elites defended democratic values and discussed polling information suggesting mass commitment to democracy across party lines. We report two main sets of findings. First, despite generous financial incentives, there was substantial differential compliance by partisanship, though not by pre-treatment attitudinal measures of affective polarization or anti-democratic attitudes. Second, our intervention achieved a durable reduction in beliefs that opponents were opposed to democracy (measured at 1 week and 2 months post-event) and in willingness to attend future depolarization events. However, we found no reduction in anti-democratic attitudes across many pre-registered outcomes. We use a follow-up survey experiment to explore the mechanisms underlying both of these findings.

Word count: 10770

[*] Authors acknowledge funding from the Polarization and Social Change Lab and thank Liz Joyner and Village Square for cooperating on the experiment. This study was approved by the Columbia Institutional Review Board, Protocol number AAAV3957. The analysis was pre-registered on OSF at https://osf.io/67ezx/files/osfstorage

[†] PhD Candidate, Department of Political Science, Columbia University

[‡] PhD Candidate, Department of Political Science, Columbia University

[§] Assistant Professor, Department of Political Science, University of Pennsylvania

1 Introduction

Organizations that seek to depolarize Americans have developed promising methods for reducing polarization and correcting harmful false beliefs about opposing partisans. For these techniques to change American public opinion at a scale sufficient to transform the incentives faced by politicians, they must reach individuals who hold dangerous, inaccurate, or hostile beliefs about political opponents. They also ought to deliver generalizable and cost-effective messages with enduring effects. In today's politics, these processes are impeded by the same mechanisms that create misperceptions in the first place: geographic (Brown and Enos, 2021) and social sorting (Mason, 2018) as well as heavily partisan mass and online media (Levendusky, 2013) and a growing partisan gap in institutional trust that results in Republicans feeling hostility towards the institutions, namely academic and non-profit organizations - at the heart of many depolarization initiatives Zhang (2023).

At the same time, elite-driven depolarization efforts are often characterized by an idealistic and anti-partisan tone, which may well damage candidate prospects in primary campaigns or call into question their partisan bonafides. Perhaps nowhere are these challenges clearer than in the case of meta-perception corrections. This style of intervention takes as a premise that American partisans have exaggerated beliefs about the opposing party's anti-democratic attitudes at the mass level (Braley et al., 2023; Pasek et al., 2022; Ahler and Sood, 2018) and that simple corrections to these beliefs can reduce support for anti-democratic attitudes. However, elite-led corrections require politicians to make positive remarks about opponents and partisans who hold exaggerated meta-perceptions to opt-in to contexts where they may hear such remarks.

To evaluate the effectiveness of an in-depth elite-led meta-perception corrections, we employ a field experiment in partnership with a depolarization organization. Our bundled intervention featured not only a meta-perception correction, but also an hour-long (of which 30 minutes were required for participants to be paid) demonstration from party elites of how to have civil discourse and and disagree on policy issues while maintaining strident political disagreement on issues of substantive concern. The event featured regular invocations of democratic norms surrounding free speech and the importance of expressing beliefs absent fear of retaliation. Our goal was to administer an intervention through a realistic discussion of the type that politicians could routinely engage in without fear of backlash from zealous co-partisans. We test the effect of this intervention on not only meta-perceptions of the other side, but also respondents' own support for anti-democratic actions, especially those relating to free speech norms and censorship. We supplement these attitudinal outcomes with several quasi-behavioral outcomes, gauging respondents' commitment to pro-democratic behaviors outside of the survey.

To investigate the practical challenges of scaling depolarizing interventions, our design allows assessing compliance among participant sub-group in order to explore preregistered hypotheses about which groups were most likely to attend a depolarization event. We investigated this mechanism by randomizing offers to attend our event among a sample of individuals whose partisanship and commitment to democracy are known ex ante through a pre-survey. In a follow-up survey, we used a placebo-controlled survey experiment to derive an additional causal estimate of differential compliance with depolarization treatments.

Ultimately, our bundled, free-flowing, unscripted treatment successfully and enduringly reduced metaperceptions that opposing partisans supported antidemocratic behaviors. These intent-to-treat effects are substantively similar across subgroups, despite partisan gaps in event attendance: Republicans were less likely to attend our treatment event, conditional on an offer being made. This result helps to explain observational patterns of differential partisan attendance at depolarization events, though no such compliance gap exists for more affectively polarized or anti-democratic subjects. Third, we find that intentions to participate in future depolarization events increase with event-offers and attendance. However, the treatment had no effect on our other attitudinal outcomes, from support for undemocratic practices to defense of free speech, nor on our other quasi-behavioral outcomes. These null results on attitudes are robust to modeling choices and pre-registered sub-groups analyses. These findings help us to conclude that attending depolarizing events can create significant and enduring reductions in negative beliefs about out-partisans, while also increasing interest in future events without increasing support for democratic values. We conclude that depolarization events can work to durably reduce negative beliefs about out-partisans, while also increasing the likelihood of future event attendance.

2 Background

Our paper is rooted substantively in a literature on depolarization initiatives and meta-perceptions corrections, as well as a set of papers that investigate the logistics of participatory interventions (Hanson et al., 2025). Drawing on qualitative accounts from practitioners and prior academic research, we identify three challenges to the practical effectiveness of depolarization initiatives: 1) The ephemeral nature of democratic meta-perceptions; 2) Identifying depolarization messaging that self-interested partisan politicians are willing to repeat at scale; and 3) Self-selection into depolarization events that renders it difficult to administer interventions to the subjects who hold harmful attitudes at baseline. We designed our field intervention to investigate all three of these pitfalls. Treatment consisted of a memorable, 30-minute long event; we used real political elite messengers who demonstrated respect for democratic norms amid issue disagreement; and we used an online sample of compensated participants. Compared to a recent experiment by (Weiss et al., 2025), we sacrifice a degree of external validity - using paid participants and administering treatment through an online platform, instead of through targetable advertising. The benefit of our approach is in providing the treatment to participants who would otherwise not have been receptive to depolarization messaging.

2.1 Meta-Perceptions

Substantively, our experiment is rooted in a literature on second-order beliefs about opposing partisan's commitment to democracy, which we refer to as *meta-perceptions*.[1] We review prior findings regarding this type of belief and describe how our multi-pronged intervention relates to simpler polling treatments. Our intervention, which includes messengers and factual corrections targeting both parties, serves to bundle a number of prior treatments that proceed along similar theoretical axes.

We draw on a literature that describes elite and mass support for democratic backsliding as rooted in dynamic beliefs about political opponents. At the elite level, formal models present competition over democratic rules as resembling an indefinitely iterated prisoner's dilemma in which fear of opponent's can constrain backsliding (Weingast, 1997; **?**; Helmke et al., 2022). Meanwhile, qualitative accounts emphasize the "tit-for-tat" nature of the erosion of democratic norms, in cases as diverse as Weimar Germany, early-2000s Venezuela and the contemporary United States Levitsky and Ziblatt (2018). Broadly, these formal and qualitative accounts suggest that beliefs about the other party's democratic commitment can impact one's own attitudes about democracy and support for pro and anti-democratic behaviors.

Recent work has extended this intuition to the mass public - and scholars have sought to shift these mechanisms as part of a broader set of interventions to reduce anti-democratic values and affective polarization (Levendusky, 2023; Wuttke and Foos, 2024; Voelkel et al., 2024). Specifically, these approaches have explored whether "meta-perception corrections" can reliably promote pro-democracy attitudes. There have been promising successes from this style of intervention, most notably the large reductions in support for anti-democratic or violent actions observed by Braley et al. (2023) and Mernyk et al. (2022)

[1] We use this term to refer to beliefs about support for anti-democratic behavior at the mass level of the opposing party, though it might in other contexts refer to broader 2nd order beliefs that encompass co-partisans. We view the parties as having distinct beliefs about democratic norms(Panizza et al., 2024)

in a mass sample and by Druckman et al. (2023) at the elite level. These papers employ "ask-tell" treatments where respondents are asked their priors about out-partisan attitudes and then randomized to a control which repeats their answers or a treatment which provides correct shares of opposing partisans side-by-side with the respondents' initial belief. Because of the systematic over-estimation of opposing partisan support for democratic backsliding[2] these corrections cause updating towards opponents being committed to democracy in the vast majority of experimental subjects.

The first obstacle to enduringly changing attitudes using this mechanism is rooted in the nature of meta-perceptions themselves as highlighted in a recent study by Dias et al. (2024). This study finds that democratic meta-perceptions are highly unstable in a panel survey. Further, these authors find inconsistent treatment effects of corrections and that more proximate attitudes are easiest to move in response to corrections; a finding that aligns with several recent null results in experimental attempts to reduce anti-democratic attitudes (Wuttke et al., 2024). Finally, Druckman et al. (2023) notes that even mild counter-arguments undermine the effectiveness of corrections. At issue in these debates is whether work on meta-perceptions constitutes a method for reducing anti-democratic attitudes or merely a description of underlying psychological processes that is difficult to alter outside of the controlled settings of a survey experiment. Our partnership with a bridging organization aims to investigate this question.

Further, despite impressive empirical results, existing scholarship has not clearly defined the cognitive mechanisms through which meta-perceptions operate. Both strategic and affective mechanisms could be in play, such that updating about the opposing party's mass support for democracy reduces negative affect (similarly to learning out-partisans oppose the party-stereotypical position on a policy issue (Orr and Huber, 2020; Orr et al., 2023)). While affective polarization does not appear to causally affect support for anti-democratic

[2] These misperceptions are a specific instance of a broader pattern of false beliefs about political opponents in the United States (Ahler and Sood, 2018) and incorrect second-order beliefs about both in and out-groups Bursztyn and Yang (2022)

behavior (Broockman et al., 2023) , related concepts may have a more robust interaction with these preferences (Finkel et al., 2024). In contrast, a purely strategic explanation for the relationship between second-order beliefs and support for democratic behaviors suggests that meta-perceptions should matter insofar as they shift concrete beliefs about how the opposing party will behave, a concept we test with “prediction” questions that assess beliefs about concrete outcomes. Finally, our intervention serves to treat beliefs about co-partisans, which could promote pro-democratic attitudes through simple conformity pressures (Valentim, 2024), though the diffuse nature of the group - nationally distributed co-partisans, might blunt the effectiveness of this portion of the intervention.

2.2 Messengers of Depolarization

Second, we aim to test a version of a depolarization treatment delivered by overtly partisan actors. While we do not randomly vary the identity of the messenger, our intervention tests an unusual form of bridging initiative that couples a standard depolarizing intervention with robust, civil policy debate. Because paid exposure to corrections is prohibitively expensive at a scale (again see Dias et al. (2024) for a discussion), messages delivered by politicians of their own accord are easier to scale, especially because prominent figures will generate earned media that can mitigate the financial challenge of scaling these messages. Further, these messages constitute a special incidence of elite opinion leadership directed at promoting pro-democratic attitudes Wuttke and Foos (2024); Wuttke et al. (2024).

This mechanism poses a challenge in the contemporary political environment. While politicians have both made many depolarizing statements and (more rarely) participated formally in depolarization initiatives Voelkel et al. (2024); Weiss et al. (2025), most notably Utah Governor Spencer Cox, (Voelkel et al., 2024), there is mounting evidence that politicians who defy their own party on democratic norms may be more likely to be electorally sanctioned(Banda and Sievert, 2024; Bartels and Carnes, 2023) and that agreement with and civility towards out-partisans brings reputational costs (Hussein and Wheeler, 2024).

Politicians willing to deliver the most idealistic version of these messages are few and far between. These constraints make idealistic depolarization initiatives challenging to scale by cross-pressuring the subset of political elites most able to generate earned media. For the sake of realism of the intervention and viability for replication in politics across the U.S., we deliver our treatment event as part of a realistic political discussion which included strident partisan disagreement. Our design requires elite actors to neither repudiate their party nor engage in the kind of rhetoric that may call their partisan credentials into dispute.

2.3 Selection into Depolarization

Third, and most significantly, we consider the challenges of selection into depolarization events at the attendee level. The meta-perception correction that forms the core of our bundled intervention is one of a set of techniques designed to reduce affective polarization or anti-democratic attitudes. From our background interviews, organizers of these events repeatedly reported ideological and demographic homogeneity in their attendees. Specifically, these initiatives reported three challenges. First, they had difficulty attracting Republicans, especially Trump-supporting Republicans. Second, attendees across party lines with elevated levels of affective polarization or anti-democratic attitudes were infrequent attendees. Third, depolarization initiatives reported a paucity of non-college educated subjects. Generally, practitioners report a sense of "preaching to the choir." While theoretically, this selection problem is common across many forms of depolarization event, the prior effectiveness of meta-perception corrections accentuates the stakes of this obstacle, since this is an intervention we reasonably expect to prove effective, conditional on compliance.

To more systematically evaluate recruitment challenges, we reviewed survey records of event participants at 19 depolarization organizations (total N of 1915 attendees). From this sample, 68.56% identified as Liberal or Lean Liberal, 11.59% as Conservative or Lean Conservative, 4.75% as neither leaning Liberal or Conservative, and 15.1% provided no information about their ideological leaning. Despite explicitly bipartisan or non-partisan organizational identities, rhetoric and event advertising, these groups recruited few conservatives. Anecdotally, depolarization activists described many of even this modest group of conservative attendees as Trump skeptics, far out of proportion with their shares of conservatives or Republicans nationally. We do not have evidence on readily comparable scales about these groups levels of affective polarization, but qualitative accounts suggest low levels of affective polarization among the staff, attendees and donors of depolarization initiatives.

Few existing studies assess this selection mechanism, in part because the mechanics of such interventions often involve either survey experiments where non-compliance is rare Voelkel et al. (2024) or experiments that condition on an explicit willingness to attend a specific event. An exception is a recent paper that explored "preaching to the choir" mechanisms among community policing efforts Hanson et al. (2025). More generally, we note wide partisan differentials in trust in expertise, higher education and a range of supposedly neutral mediating institutions (Zhang, 2023), a trend that extends to explicitly non-political arenas (O'Brian and Kent, 2025).

Crucially, our design does not fully solve this challenge as it selects for participants of Cloud Research Connect who were willing to answer a survey with political questions - and who presumably place a greater weight on financial incentives than the broader public. However, our sample does include large sub-samples from both parties and a wide range of (pre-treatment) affective polarization ratings and democratic attitudes. Our sample includes considerable numbers of different political blocks, even Trump-supporting Republicans normally unreachable by depolarization initiatives.[3] By observing partisanship and democratic attitudes before event recruitment we can isolate whether these characteristics correlate with event attendance conditional on treatment assignment. In contrast, most existing studies of attendance explore either responses from attendees, which serves to select on the dependent variable of attendance, or use survey based measures of willingness to attend events instead

[3] ??? of the Republicans in our sample said that they believed Joe Biden did not legitimately win the 2020 presidential election, somewhat lower than the approximately 68% of Republicans nationally

of behavioral evidence of actual event attendance.

3 Experimental Design

To test both the effectiveness of our bundled elite-driven depolarization effort and explore the broader practical challenges faced by depolarization initiatives, we carried out a field experiment on an online sample using compensated offers to attend a depolarization event and recorded our main outcomes through a follow-up survey offered one week after the main event. The experiment was designed to test externally valid and scalable forms of elite speech by employing a treatment that partisan politicians would willingly deliver to their constituents. To further investigate persistence, differential attendance and the mechanisms of meta-perceptions corrections, we conducted a follow-up survey experiment on the same sample (N=2181 responders 10 weeks after the treatment event).

3.1 Randomization and Procedure

We recruited participants for our study on the online survey platform Cloud Research Connect. A screening survey collected demographic information, initial attitudinal measures, and respondents' availability to attend a future event. We asked respondents what party they identify with or, in the case of independents, lean towards. We excluded from the study "pure independent" respondents who reported not having any partisan leaning. We identified a sample of 3,461 eligible respondents (1,912 Democrats and 1,564 Republicans), which we block-randomized into treatment (N = 1,730) and control (N = 1,731) by party * region block. Geographic regions consisted of the four time zones of the contiguous United States plus a separate block for Florida respondents (with the small number of Alaska and Hawaii residents included in the Pacific time block). In the table below, we show that treatment and control groups display balance across age, gender and education levels as well as pre-treatment survey measures of affective polarization and democratic meta-perceptions. The

	Control Group		Treatment Group			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Age	41.4	13.4	41.0	13.6	-0.4	0.5
Aff. Pol.	51.7	29.8	51.9	29.4	0.2	1.0
Meta-Perceptions	25.6	27.2	25.6	27.5	0.1	1.0
Male	0.4	0.5	0.4	0.5	0.0	0.0
College Educ.	0.6	0.5	0.6	0.5	0.0	0.0

Table 1: Treatment and Control Groups Comparable Across Demographic and Attitudinal Covariates

party-region block randomization was designed to ensure that within each block, Democratic and Republican partisans were exposed to an offer with a comparably (un)pleasant time of day to attend the event, avoiding a correlation between partisanship and geography that would complicate our pre-registered compliance analysis.

Below, we show demographic balance between (assigned) treatment and control groups. Of note the sample, was more female and college educated than the general public.

3.2 Treatment Event

One week after the screening survey, we invited the treatment group to attend the treatment event on September 18th, 2025. The invitation described the event as a “Bridging Partisan Divides Event.” Respondents were offered \$10 to attend the event for 30 minutes[4] and complete two attention checks. The treatment group received repeated reminders about the timing of the event. The event was held on a Thursday at 7:00 PM US Eastern time. We closed the event to new participants 10 minutes after it began. The day after the event, we offered all treatment group users who had not attended the live event a new opportunity to watch a recording of the event. Respondents who watched the recorded event received the same attention check questions and compensation.

[4] This price was confirmed to be a very generous offer in the context of the usual rates for Cloud Research Connect tasks

The treatment event was a conversation between a Republican state representative and a Democratic local religious leader. The nature of this event was more free-flowing between both parties, with the explicit intention on showing civil discourse between a Democrat and Republican. As such, the treatment here is not limited to any particular stimulus, but is rather the bundle of the entire conversation. Early in the event, we asked participants to mentally guess about the correct answers to our meta-perceptions questions. We then provided respondents with the correct figures drawn from existing polling and our pre-treatment surveys. These factual reminders - approximating a survey experimental ask-tell intervention - served as the starting point for a longer discussion about respect and listening across party lines, A main takeaway from the event was that it framed the speakers' life stories and friendship despite their partisan divide. The speakers also discussed their shared commitment to American democracy and the importance of defending the freedom of speech of members of the opposing party. Our event thus featured political elites modeling civil behavior and adherence to democratic norms even while (at times) disagreeing on policy issues. Following the 30 minute mark (when participants were able to leave and receive their full payment), the discussion turned to discussing a series of individual political issues, including immigration.

The 89% correct responses to the attention check questions[5] suggests our respondents were overwhelmingly attentive, though we do not condition on successful attention check completion due to the post-treatment nature of this variable. And although their attendance was paid, we find evidence that many respondents found the event enjoyable and engaging. Respondents were paid to attend for 30 minutes, but had the choice to keep watching after that required period. About 15% of compliers (those who watched for 30 minutes) watched for at least 45 minutes[6], and 8% for at least 60 minutes. This "super-compliance"

[5] Attention checks were the country of origin and state of residence of the two guest speakers, details that were not obvious from the title of the event but were abundantly clear to anyone who paid attention. Respondents knew that there would be an attention check, but did not know the questions until after they finished watching.

[6] Though we acknowledge the possibility that these respondents left the event playing, their demographic

was particularly common among Black respondents - of note, the two guest speakers at the treatment event were both Black political leaders from Florida. Following the conclusion of the event, we received messages from respondents expressing their enjoyment of the event.

We consider the event in the context of prior efforts to correct inaccurate meta-perceptions about out-groups, in this case opposing partisans.[7] (Moore-Berg and Hameiri, 2024) provide a typology of how meta-perception corrections can be operationalized. Our treatment in their framework is a "framed" treatment in that it provides factual information embedded in a narrative of bipartisan embrace of democracy to aid interpretation. We also combine features of direct and indirect interventions - giving both correct factual information and allowing treated subjects to view evidence of an opposing partisan showing their commitment to democracy, a form of "parasocial contact" (Moore-Berg and Hameiri, 2024) that demonstrates that opponents are behaviorally committed to norms of open discourse. The extended length of the event was also designed to deliver a stronger version of existing treatments.

3.3 Outcome surveys

We launched the main outcome survey 8 days after the event and 6 days after after we closed the recorded event offer. The main outcome survey was open for the next 3 days, and an additional opportunity to complete the survey was kept open until October 4th, such that respondents could take the outcome survey between 8 and 15 days after the event was held. Notably, we do not find heterogeneous effects between respondents who answered the outcome survey early versus late.

To mitigate demand effects, we ensured that the outcome survey had no overt con-

characteristics leads us to believe at least some of these respondents watched the full event intentionally

[7] Though as discussed in our literature review, we note the effort to simultaneously treat both parties serves to bundle our treatment with a correction about co-partisans, which could affect social norms among an in-group, though we would expect this mechanism to further contribute to reductions in expressed anti-democratic attitudes

nection with the pre-treatment screener survey or the treatment event[8]. While the mechanics of the intervention required that both treatment and surveys were administered through the same platform, we endeavored to break the link between these tasks. Because respondents likely took many other political surveys during this time period, they would not necessarily link the treatment event to the outcome survey. In addition, the gap between surveys, the likely exposure to other political surveys (the survey occurred during the 2024 presidential election campaign, and Cloud Research hosted many political surveys) and recent research on repeated measures (Jordan et al., 2025), suggesting that there is little risk of consistency bias, especially compared to the power gains from a highly prognostic pre-treatment covariate. We achieved a recontact rate of 92% in the control group and 93.5% in the treatment group, a difference that is not statistically significant[9]. We discuss the distribution of non-responses and how this may impact our estimates in the attrition section below.

3.4 Outcome measures

We measured a wide range of survey experimental and quasi-behavioral outcomes to test for changes in both attitudes and real-world behaviors. We list our outcome measures of interest in Table 2. The complete survey is available in the Appendix.

[8] The label on the event offer and account name with which the task was associated differed between these tasks on the recruitment platform

[9] The specific re-contact procedures are available in the appendix

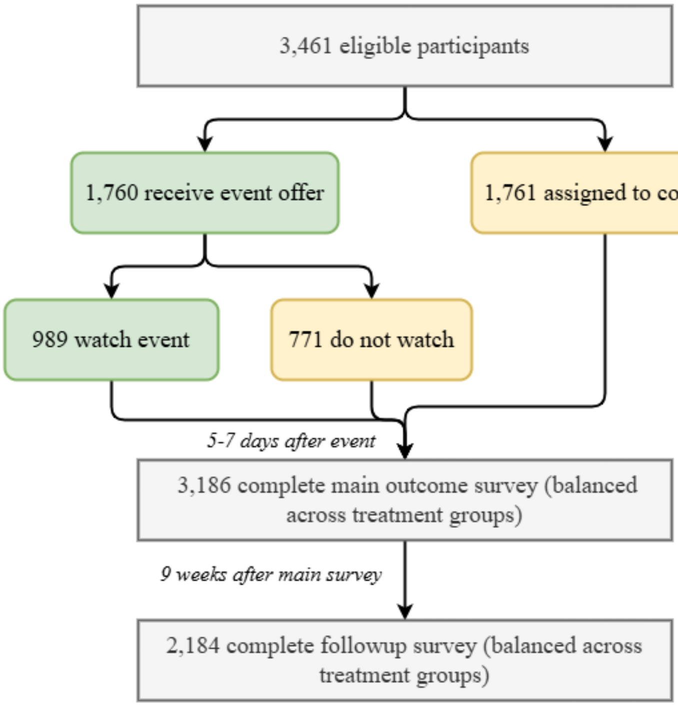
3,461 eligible participants
1,760 receive event offer
1,761 assigned to control
989 watch event
771 do not watch
5-7 days after event
3,186 complete main outcome survey (balanced across treatment groups)
9 weeks after main survey
2,184 complete followup survey (balanced across treatment groups)

Figure 1: Experimental design

Drawing on the findings from Dias et al. (2024) we identify some measures as more clearly proximate to the correction and therefore requiring less complex strategic reasoning for a correction to affect these beliefs. Our questions draw from a combination of existing survey measures common to the literature on democratic norms and customized questions designed to elicit attitudes specifically addressed in our intervention.

We measure meta-perceptions as the percentage of the other party's voters subjects believed to hold the belief in question, from 0 to 100. Our attitudinal questions had Likert responses that we converted to numeric scales, with strongly disagree equal to 1 and strongly agree equal to 5. Results are substantively unchanged under alternative codings (such as a binary agree vs disagree outcomes).

We chose our undemocratic practices support questions to have variation in the distribution of responses and thus mitigate ceiling effects. If, for instance, 95% of respondents opposed the undemocratic practice at baseline, there would be no room for a meaningful

Outcome	Questions
Meta-perceptions battery	Average of four questions asking what percentage of opposite party members agree with the following:[10] • It is justified for [Opposite party] to use violence if the [Own Party] is declared the winner of this presidential election. • Supporters of the [Opposite party] should harass members of the [Own Party] online so that they feel frightened or afraid they might lose their jobs. • The government should be able to censor media sources that spend more time attacking [Opposite party] than [Own party] candidates. • Members of the [Opposite party] should use violence against peaceful protests organized by members of [Own party].
Predictions of post-election anti-democratic behaviors	Average likelihood of the following actions happening after the 2024 election: • The [Opposite party] will try to silence media outlets who support [Own party] by changing the rules to make it easier to sue them. • The [Opposite party] will use violence to try to silence protesters who belong to the [Own party]. • The [Opposite party] will have leading [Own party] arrested without evidence.
Support for undemocratic practices	The same as the meta-perceptions battery, with own party and opposite party reversed
Affective polarization	Difference in responses on 100-point feeling thermometer between own and opposite party.
Quasi-behavioral outcomes	The final section of the survey gave participants the option of taking the following actions: • Requesting information about how to sign up to work the polls during the 2024 election • Signing a public pledge to not discriminate against others based on their political beliefs. *Respondents were informed that their names and state of residence would be made publicly available under the pledge.*[11] • Signing the "Team Democracy" pledge expressing support for American democracy • Signing up to be recruited for future events by an organization that seeks to bring Americans together across party lines
Support for undemocratic election tactics (Secondary Outcome)	Average support for own party taking the following actions during the 2024 election: • Close polling places in areas that support the other party • Spreading lies about political opponents • Not accepting the results of the election if your candidate loses
Defense of free speech (Secondary Outcome)	Average agreement with the following: • I have a responsibility to defend [Opposite party members] I know when [Own party members] attack them for their beliefs • The law should protect members of the [Opposite party], even if they spread lies about [Own party] candidates. • The government should continue providing licenses for channels like [FOX News / MSNBC] which give biased coverage in favor of the [Opposite party].

Table 2: Primary outcome questions

treatment effect. Control group responses show us the ceiling in our potential treatment effects. We find 61% of control group respondents answered “strongly disagree” to all four undemocratic practices prompts, and 75% answered “strongly disagree” for each question on average. These results suggest that ceiling (or rather floor) effects could have blocked movement in anti-democratic attitudes, since many respondents were maximally opposed to undemocratic behaviors in the control group. Nonetheless, there was ample space for a reduction in anti-democratic attitudes across all measures, and the behavioral measures had no such limitations.

3.5 Follow-Up Experiment

In addition to our main experiment and survey outcome measures, we conducted a follow-up two months after the main event. This survey was available to all subjects who participated in the initial screener survey and re-asked our meta-perceptions battery was also including additional survey experiments to explore differential compliance by partisanship and the mechanisms through which meta-perceptions operate.

4 Results

We report the results of the main experimental intervention in accordance with our preanalysis plan. We first discuss our first stage, compliance results and then our main outcome estimates before examining heterogeneous effects for both preregistered and exploratory analyses. Our models are specified as follows where Y_i is the outcome variable, χ is a vector of preregistered covariates: education, race, partisanship, presidential vote choice, pretreatment attitudinal variables, and τ is a vector of block-party dummy variables.

$$Y_i = \beta_{Z_i} + \omega\chi_i + \Omega\tau + \epsilon_i \tag{1}$$

4.1 Compliance

We first report compliance results. We note that this first-stage outcome is observed without attrition because failing to "take" the treatment is observable by the absence of a subject's ID from the pool that participated in the task. Our pre-registration defined compliance as attending the event for the full 30 minutes required.

We find an overall compliance rate of 59% (1,019/1,730) with significant differential compliance by partisanship.[12] Among Republicans, there were 390 compliers among 691 assigned to treatment (56.4%). Among Democrats, there were 609 compliers out of 903 assigned to treatment (67.4%). Regression results for the first stage outcome are below. We note these subgroup effects do not have a causal interpretation and partisans may (and do) differ on other characteristics besides their partisan identity. The partisan gap in attendance was directionally the same across all 5 geographic blocks (see Appendix Table **??**).

Practitioners have noted the difficulty of convincing Republicans to attend depolarization events. Knowing this, we structured our treatment event to be appealing to Republicans through including a Republican politician giving depolarization messaging. Despite this, we *still* observed a substantial gap in compliance.

Next, we explore how compliance is predicted by two pre-treatment covariates in more granular ways. In exploratory conversations, several depolarization practitioners expressed concerns that only individuals with low pre-treatment levels of affective polarization or low meta-perceptions would be willing to participate in depolarization initiatives. While we acknowledge the external validity challenges of generalizing from our online sample, we find little evidence that paid participation is predicted by these attitudinal measures. In Figures 2 and 3 below, we show that there a little evidence of a strong relationship between

[12] To preserve the ecological validity and verisimilitude of the treatment event, our partner organization's website did describe the event and include a link to the livestream. The partner organization did not, however, widely advertise the event. We estimate that approximately 50-100 people unconnected with the study attended the event. While hypothetically possible, it is extremely unlikely that any of the members of the control group attended the event in this way. We thus treat this as a case of one-sided noncompliance.

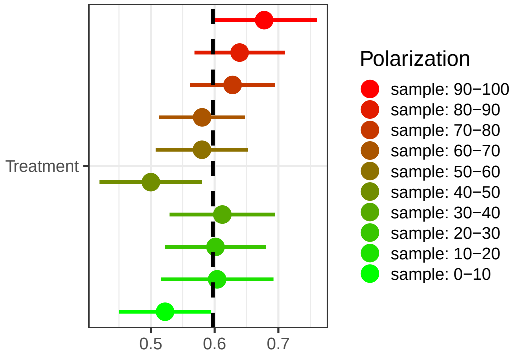

Figure 2: Affective Polarization

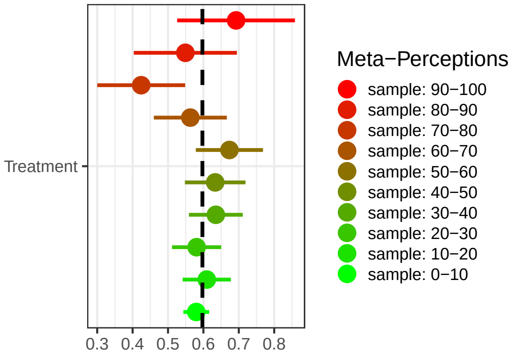

Figure 3: Meta-Perceptions

either attitude and compliance. In these figures, the vertical line is at the average compliance rate of 59%. Directionally, more affectively polarized individuals are likelier to attend the event, as shown in full regression models in the appendix, though this difference does not approach statistic significance in continuous or binner models.

In Figure 10 (appendix), we show continuous versions of these models, demonstrating that these pre-treatment measures do not significantly predict the effectiveness of our financial offer at encouraging actual event attendance. Directionally, more affectively polarized individuals are likelier to attend (and higher affective polarization significantly predicts accepting our offer, at least in some model specifications) and individuals with lower meta-perceptions are likelier to attend.

In our study, there is some discretion in how to define which subjects received, or complied with, the treatment. Above, we have used our pre-registered definition that any respondent who watched the treatment event for the required length of time (30 minutes) was a complier. We could alternatively define compliance as watching and answering both attention check questions correctly, which would reduce compliance to 55.7%, or even as watching the event live and not counting respondents who watched the event recorded, which would reduce compliance to 27.8%. These alternative definitions of compliance do not substantively change the aforementioned patterns seen in compliance rates.

As a note on external validity of our compliance estimates, our experiment is likely to over-state compliance compared to other contexts. This is for two reasons. First, the sample conditioned on having an account on Cloud Research Connect, a website in which participants take surveys and complete other tasks for small payments. Second, our event was well compensated by the standards of paid online tasks. As such, we can interpret non-compliance with treatment as demonstrating a particular disinterest in depolarization events. In a real-world context, compliance is likely to be substantially lower across groups, which could lead to substantively larger partisan divides in willingness to attend depolarization events.

4.2 Main Results

We begin by estimating our main results, reporting both intent-to-treat and complier average causal results. We report our results first for the four main pre-registered survey batteries. Because of the specific structure of the intervention, the meaning of the ITT is the effect of an incentivized offer to attend the event, while the CACE is the effect of the offer on those who choose to take treatment by watching the event.[13] All models include our pre-registered vector of covariates including demographic controls, attitudinal measures from the recruitment survey, and dummy variable for blocks.

First, we test whether the meta-perceptions correction in our treatment event had an enduring effect on meta-perceptions several days later. This is a much harder test of a meta-perceptions correction than a correction and test in a standalone survey experiment. Not only was there a multiple day delay (between 5 and 10 days) between the treatment and the outcome survey, but the outcome survey was not overtly connected to the treatment event. As users of a survey-taking website, respondents presumably answered other political surveys in the meantime and would not necessarily link the outcome survey to the event that

[13] While paid attendance is rare for depolarization events, many organizations offer compensation (for example, by raffling gift cards) for attending events or filling out associated surveys.

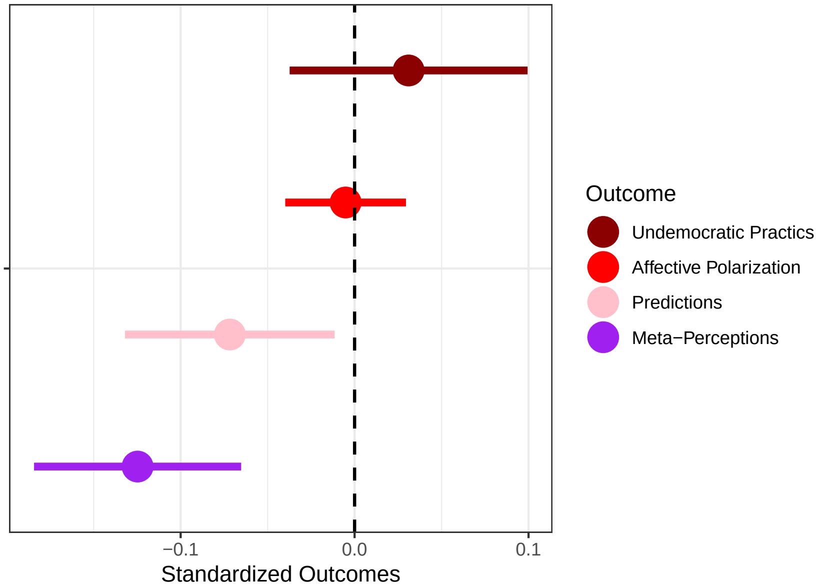

Figure 4: ITT Effects on Survey Outcomes

they had attended several days earlier. We did not inform event attendees that there would be an outcome survey afterwards, in order to mitigate demand effects.

We present our estimated ITT effects on attitudinal outcomes in Figure 4. We find a modest but highly significant ITT effect on meta-perceptions of 0.12 standard deviations ($p < 0.001$). We do not find a measurable decline in treatment effects between participants who completed the survey early and those who completed it late. Treated participants thus learned from the meta-perceptions correction and remembered it for at least 10 days. We find significant treatment effects on all four meta-perception questions, with the largest effect on the "[Opposite party] should use violence against peaceful protests by [Own party]" question. We did not pre-specify hypotheses about which individual questions would see the greatest treatment effects. We also find a negative treatment effect on predictions of postelection anti-democratic behavior ($p < 0.05$). That is, the treatment made respondents more optimistic that the opposing party *would not* violate democratic norms after the election. In a later section, we will consider important heterogeneity in this effect.

However, we find *no treatment effect* on our question batteries about political attitudes, as opposed to beliefs about the opposing party. The treatment did not change respondents' support for undemocratic practices (both censorship and unfair electoral practices), support for the opposite party's free speech rights, or affective polarization. These null effects are before any correction for multiple comparisons and point estimates are substantively small in addition to statistically insignificant. In the appendix, we report the results on two additional sets of outcomes, showing further null effects on attitudes.

The discordance between the observed effects on beliefs about the other party and the lack of effects on support for undemocratic practices is even more striking given the survey design. The outcome survey presented the support for undemocratic practices questions immediately after (though not on the same page as) the meta-perception battery. This survey design thus implicitly encouraged respondents to think reciprocally about the out-party's beliefs and their own. Further, one of our question batteries on which we observed no significant movement was identical to the meta-perception battery that was the subject of treatment (with party names swapped), making it exactly the type of proximate-to-treatment outcome most likely to be shifted by altering meta-perceptions in prior literature (Dias et al. (2024)). As noted in our pre-analysis plan, these analyses were well-powered to detect substantively small treatment effects on the order of $\frac{1}{10}$ of a standard deviation.

Did our treatment and its effects on meta-perceptions translate into changes in democracy-related behaviors? In Figure 5, we show ITT effects for our four behavioral measures and a combined index of the four. As before, all models include our pre-registered vector of covariates. The treatment had a null effect on the combined index, but with considerable variation between individual behaviors. We find a large and highly significant treatment effect on respondents' interest in attending a future depolarization event. In other words, the treatment event made respondents more interested in attending similar future events, but had no direct effect on pro-democratic behaviors.

4.3 Complier Average Causal Effects

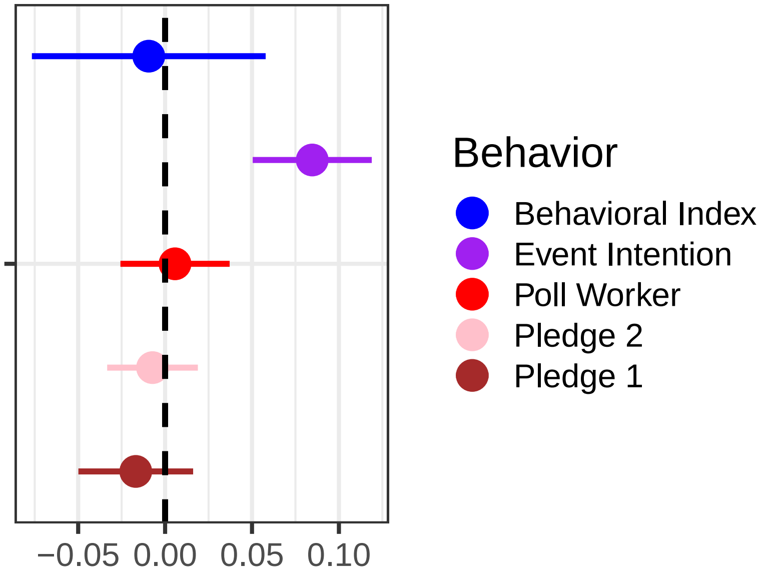

Figure 5: ITT Effects on Binary Behavioral Outcomes

Table 3: Comparing CACE and ITT estimates

	Metas	Preds	Aff. Pol.	Undem Practices	Behaviors index
ITT Estimate	−0.125***	−0.072*	−0.005	0.031	0.016
	(0.030)	(0.031)	(0.018)	(0.035)	(0.011)
CACE Estimate	−0.199***	−0.114*	−0.008	0.050	0.026
	(0.048)	(0.049)	(0.028)	(0.056)	(0.017)
Num.Obs.	3180	3180	3180	3180	3155
R2	0.274	0.255	0.753	0.036	0.023

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Next, we estimate complier average causal effects (CACEs) for our main outcome measures.[14], our assumption is that the control subjects who would have complied with the treatment assignment also attrited at comparable rates. We calculated CACE estimates in R with a two-stage least squares model using the ivrobust function in the estimatr package. Because we do not employ a placebo, the CACE estimates serves to mechanically re-scale

[14] We note that correlation between attrition and non-compliance complicates these estimates. While we do not observe compliance in the control group, due to the absence of a placebo. Because we observed - and expected ex ante - that compliance would be above 50%, a placebo would not have produced gains in statistical power

our ITT estimates by the amount of compliance in the population.

Notably, while there are several possible definitions of compliance, these produce small changes to the estimate and our results are robust to these alternative specifications. With our loose definition of compliance, we estimate the CACE on meta-perceptions as -0.2 standard deviations. With a stricter definition of compliance as watching the event and answering attention checks correctly, this rises to -0.21 standard deviations. The different specifications of compliance did not change which effects are statistically significant, nor did they cause substantively large changes to CACE estimates. In Table **??** below, we show the ITT and CACE estimates for the four main survey outcomes

4.4 Heterogeneous Effects

We begin by reporting results for three pre-registered heterogeneous effects: by party, by pre-treatment support for undemocratic practices and by higher meta-perceptions. Across these outcomes. Because we had also pre-registered differential compliance, these CATEs were pre-registered for CACEs rather than ITTs. For parsimony's sake, we summarize these results visually for partisan heterogeneity and through regression plots otherwise.

We test for heterogeneous effects in our estimates across two pre-treatment attitudinal variables: partisanship and pre-treatment attitudinal measures. Further heterogeneous effects models are included in the appendix. Unless otherwise specified, all models contain the same vector of covariates as the main analyses. We estimate conditional average treatment effects (CATEs) with treatment-by-covariate interaction terms.

Our treatment event provided respondents with correct statistics about opposite party beliefs, bundled with messages relating to civility and misperceptions in general. If these specific figures, as opposed to the broader message of the event, drove our reduction in meta-perceptions of opposing partisan support for democratic backsliding, we might expect to see treatment effect heterogeneity by how far respondent's priors were from the truth. In

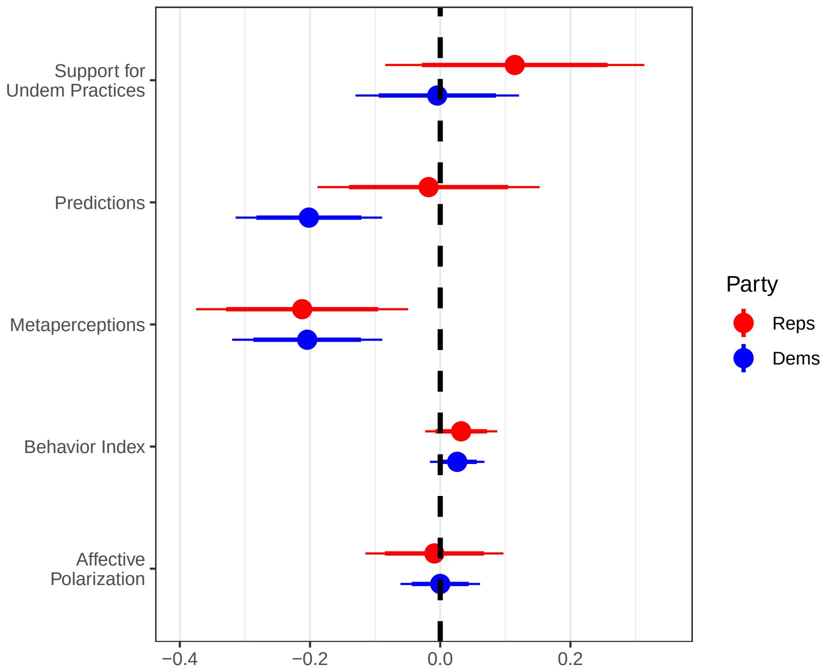

Figure 6: CACEs by Party for Main Outcomes

Table 4: Treatment Effects by Standardized Pre-Treatment Meta-Perceptions

	Metas	Predictions	Aff Pol	SUP	Behaviors
Treated	−0.125***	−0.072*	−0.005	0.031	0.016
	(0.030)	(0.031)	(0.018)	(0.035)	(0.011)
Meta-Perceptions	0.491***	0.339***	0.015	0.094***	−0.006
	(0.027)	(0.022)	(0.015)	(0.026)	(0.008)
Treated: Meta	−0.031	−0.069*	0.009	−0.035	0.003
	(0.037)	(0.030)	(0.019)	(0.037)	(0.011)
Num.Obs.	3180	3180	3180	3180	3155
R2	0.272	0.254	0.753	0.035	0.021

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Second model includes demographic covariates

Table 5: Treatment Effects by Pre-Treatment Support for Un-Democratic Practices

	Metas	Predictions	Aff. Pol.	SUP	Behaviors
Treated	−0.126***	−0.075*	−0.005	0.015	0.016
	(0.030)	(0.031)	(0.018)	(0.032)	(0.011)
Pre-Treatment SUP	0.039+	0.079***	−0.001	0.391***	0.005
	(0.022)	(0.022)	(0.014)	(0.042)	(0.008)
Treated:SUP	−0.013	−0.024	0.009	−0.010	−0.004
	(0.031)	(0.031)	(0.020)	(0.061)	(0.011)
Num.Obs.	3180	3180	3180	3180	3155
R2	0.273	0.257	0.753	0.180	0.021

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Second model includes demographic covariates

Table **??**, we see that this result holds directionally but heterogeneity is substantively modest and is not statistically significant. A one standard deviation increase in pre-treatment meta-perceptions increases the magnitude of the correction effect by 0.31 standard deviations.

we see little evidence for such a result. Instead, the treatment appears to operate uniformly across both types of pre-treatment survey measure. In the same figure, we also show that affective polarization does not predict treatment effectiveness. Regression tables of heterogeneous effects are included in the appendix. Each 1 point increase in pre-treatment average meta-perceptions increasing the effectiveness of the intervention by 0.035 points, though this affect is not statistically significant ($p = 0.25$) and is substantively small such that movement along the inter-quartile range of pre-treatment meta-perceptions changes the modeled treatment effect from 2.5% to 5%.

4.5 Persistence of Meta-Perceptions Correction

How long did the treatment event's effect on meta-perceptions endure? While our initial outcome survey was already distanced from treatment compared to a one-shot survey experiment, we wanted to test for even longer-term persistence. Our follow-up survey, conducted two months after the treatment event, began with a re-measurement of respondent meta-

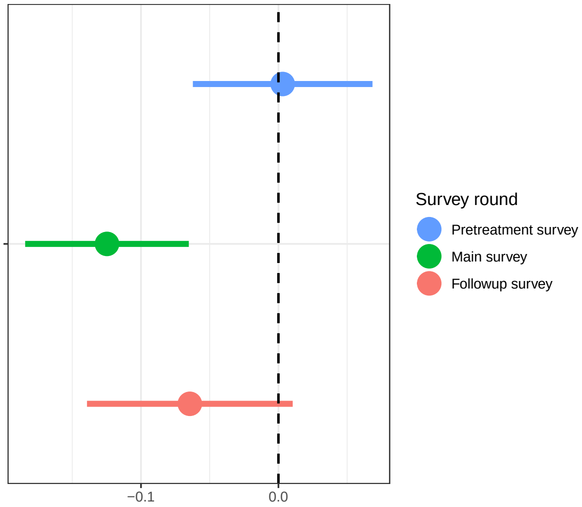

Figure 7: Persistence of meta-perceptions correction

perceptions. We used three of the four meta-perceptions questions asked in the main outcome survey, but dropped the question about support for violence in the 2024 election *because the election had already happened.* Figure 7 shows the intent-to-treat effect of assignment to the treatment group on the meta-perceptions battery collected during the followup survey. For reference, we also show the treatment effect estimated in the main outcome survey as well as a placebo test for an effect on pre-treatment meta-perceptions.

Two months post-treatment, we find a treatment effect of -0.064 SDs, which is borderline statistically significant (one-sided p value 0.046, two-sided p value 0.091). While the magnitude of the intervention's effect declines between the two outcome surveys, the effect remains detectable two months after administration of the main pre-registered outcome. This persistence is yet more striking when considering that the intervening time period between surveys included the most intense period of the 2024 election campaign and the election itself. Our follow-up suggests that attending our treatment event *durably* updated attendees' meta-perceptions across a time period with nonstop political news coverage competing for respondents' memory and attention. We find no evidence of a difference in effects by party ID (though we note that it is difficult to test for a heterogeneous effect when the main effect is already only borderline significant). The followup survey also asked a new pair of meta-perceptions question (prior to the followup survey experiment) about the opposite party's support for undemocratic practices in the aftermath of the 2024 election. We *do not* find a persistent effect of the treatment event on these novel questions. This suggests that while respondents durably remembered the corrections from the treatment event, they did not durably update their broader attitudes towards members of the opposite party.

4.6 Attrition

Our design faced a risk of attrition common to similar experimental designs (Lo et al., 2024). Especially because practical considerations that rendered a placebo infeasible,[15] we were cognizant of the risk that attrition - that is failing to complete the outcome survey - would be causally affected by the treatment, which involved an offer of compensation to attend our treatment event. To minimize this concern, we made the payment for the outcome measure sufficiently attractive that subjects across groups would be motivated to complete the survey. We repeatedly contacted respondents reminding them to complete the survey.

Ultimately, only 8% of participants failed to complete the outcome survey. Crucially, treatment assignment *did not* predict completion of the outcome survey. Treatment compliers exhibited lower attrition than non-compliers in the treatment group, though attrition was balanced overall across treatment and control conditions. The intuition behind plausible treatment-induced differential attrition is that subjects drop out of the Cloud Research Connect subject pool and some would-be drop outs might have been induced to remain active survey takers by a pleasant and well compensated experience with our event. Despite these ex ante concerns, there is no evidence of a strong relationship between treatment and reporting status. Our reporting results suggest that online recruitment platforms paired with generous incentives can achieve low and balanced rates of attrition, as offers are sufficiently

[15] Both as a question of academic interest and as a practical goal of our partner organization, we wanted to observe differential compliance with depolarization invitations by partisanship. A placebo design would depend on both conditions identifying the same set of compliers, which would have required disguising the nature of the intervention.

Table 6: Attrition is Not Predicted by Treatment Assignment

	Survey Non-Response
Assignment	−0.012 (0.012)
Num.Obs.	3496
R2	0.025
Std.Errors	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include demographic covariates

attractive to limit attrition among most respondents and there is minimal (though non-zero) churn among the participant pool over short periods.

Non-response was far higher in the followup survey than in the initial outcomes survey (37% compared to 8%), as many respondents had stopped using the survey platform in the intervening time-span. Is the followup survey's sample comparable to the main outcome survey? Non-response continued to be balanced between treatment groups.

Further, attrition is modestly predicted by partisanship, with Republicans being 6% less likely to complete the followup measure (though there is no significant interaction in models that interact treatment and partisanship to predict attrition). One explanation is that Republicans were more likely to periodically leave the platform. The partisan differential in attrition is noticeably smaller than the differential in treatment uptake.

5 Follow-Up Experiment

Our field experiment has suggested that more flexible interventions than standard survey based ask-tell corrections can durably shift meta-perceptions, but may have limited impact on democratic attitudes and behaviors. We conducted a follow-up survey experiment - embedded in our final panel wave - to assess the persistence of treatment effects and also explore why shifting meta-perceptions did not, at least in this case, changes preferences for anti-democratic behavior. We offered the followup survey to the same sample of CloudResearch Connect users that participated in the experimental screener. As with the main outcome survey, there was no overt link between the treatment event and the followup survey. We fielded the followup survey in December 2024, about 10 weeks after the treatment event and 3 weeks after the 2024 presidential election. This delay allows us to measure meta-perceptions and support for undemocratic practices in a tense post-election period when public discourse often expressed fear for the future of American democracy - especially among Democratic partisan media. Following the re-measurement of respondent meta-perceptions (described above), the follow-up survey administered an embedded survey experiment correcting misperceptions of undemocratic behaviors by either *opposite party voters* or *opposite party leaders.* This final experiment also used a placebo-controlled design to further investigate differential partisan interest in depolarization initiatives.[16]

5.1 Survey experiment: Correcting misperceptions about voters vs. about elites

Because the follow-up survey took place after the 2024 presidential election, *it would not be realistic to maintain question parallelism between Democrats and Republicans.* Democrats were in power but expecting to lose power, while Republicans were in opposition but expecting to gain power. In their different positions, the parties were capable of qualitatively different violations of democratic norms. Question wordings thus vary significantly by the respondent's political party, and we do not pool results across parties.

Table 7 includes the wordings of treatments and outcome wordings shown to respondents of each party.[17]

Respondents of each party were randomized to one of three informational con-

[16] Attrition does not affect the embedded survey experiments, as these had independent randomizations that occurred after selection into the followup survey

[17] We continue to categorize respondents by the party ID they provided in the screener survey before the treatment event or either of the outcome surveys.

Table 7: Followup survey experiment design, by respondent party

	Treatment Conditions	
Condition	**Democratic respondents**	**Republican respondents**
Shared control	As you are probably aware, the presidential election occurred last month on November 5.	
Elite correction	During Trump's first term as president, his administration obeyed the law and followed rulings of courts, even when those rulings went against his administration.	Kamala Harris called Donald Trump to congratulate him on his victory, and President Biden invited Trump to the White House to plan a smooth transition of power.
Voter correction	A recent poll found that 80% of Republican party voters believe that presidents must always obey the laws and the courts (World Justice Project, September 17).	A recent poll found that 89% of Democratic party voters accept that Trump won the election legitimately (Ipsos/Reuters, November 8).
	Outcome Measures	
Outcome	**Democratic respondents**	**Republican respondents**
Predictions of opponent behavior	• Trump will arrest Democratic politicians without evidence. • Trump will use violence to try to silence protesters who belong to the Democratic Party. • Trump will try to shut down media outlets who are critical of his administration.	• Democrats will use violence to disrupt the certification of the election. • Democrats will make false accusations in court to try to overturn the election. • Democratic states will violate laws passed by the Trump administration.
Opponent fairness	Do you think Republicans contest elections fairly?	Do you think Democrats contest elections fairly?
Support for norms violation	While Donald Trump has been declared the winner, some Democrats have suggested trying to prevent Trump from taking office. Would you approve of this?	After Trump takes office, do you think the Justice Department should pursue criminal charges against Biden administration officials for corruption, abuse of power, and treason?
Relative blame	Do you agree with the following statement: "[opposing party] politicians generally want to break the rules in American politics today?"	

ditions. The **voter correction** condition provided respondents with recent polling data showing that opposite party voters broadly support the rule of law. This is theoretically comparable to the meta-perceptions corrections in the main treatment event, although these survey results were more recent and directly relevant to concerns after the 2024 election. In contrast, the **elite correction** condition provided respondents with sourced evidence suggesting that the opposite party's leadership would respect the rule of law. The opponent elite for Republican respondents was Joe Biden, versus Donald Trump for Democratic respondents. In all of these conditions, we cite a reliable source to support the correction. We test these treatment conditions against a control condition that did not provide new information.

After the informational treatment, respondents answered four groups of outcome questions. Again, question wordings varied substantially by the respondent's party ID, as detailed in Table 7. Our outcome measures are:

We test the effects of these treatments on four outcomes.

1. An index of three questions predicting future undemocratic behavior by opposing party (*predictions)*
2. How much the respondent blames opposite party leaders relative to voters for that party's unethical behavior (positive values mean blame leaders more)
3. Whether the respondent agrees that the opposite party contested the 2024 elections fairly
4. Whether the respondent supports *their own party* taking an action that violates democratic norms. For Democratic respondents, the violation was trying to prevent the certification of the 2024 election results. For Republican respondents, the violation was prosecuting former Biden administration officials once Trump takes office. We find considerable baseline support for these norms violations (59% for Republican re-

spondents, 28% for Democratic respondents), obviating the concern of a ceiling effect masking results.

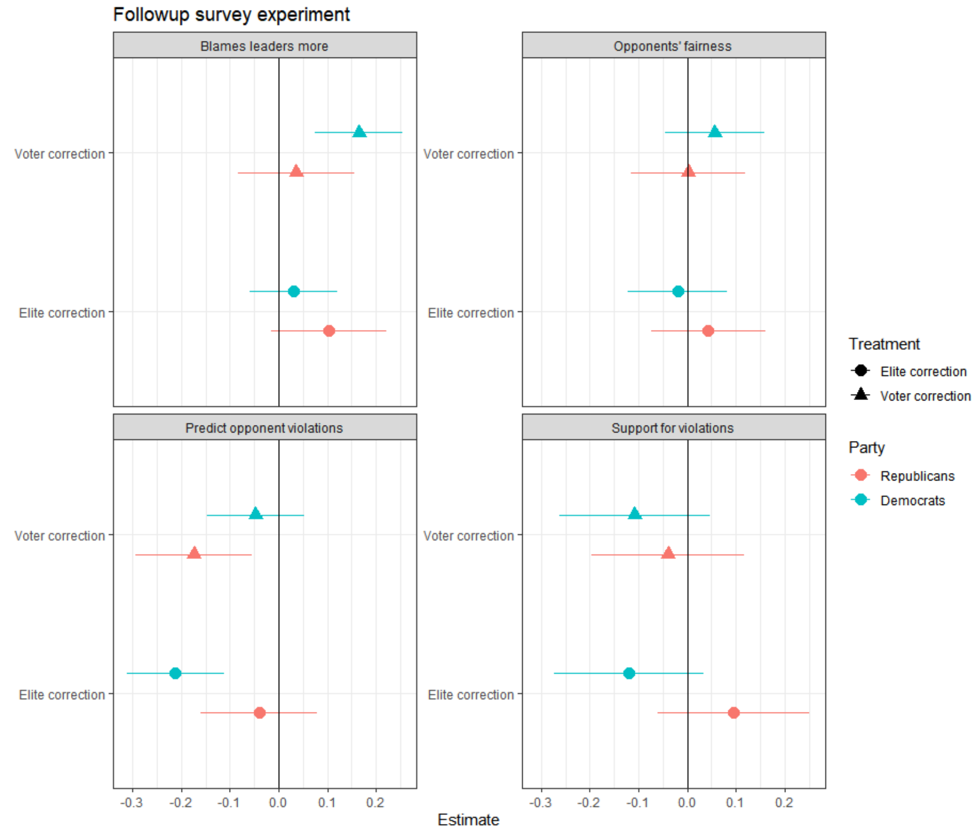

Figure 8: Effects of Elite and Voter Corrections on Main Outcomes

Figure 8shows the treatment effects of both corrections (relative to the no-information control), broken down by respondent party. Broadly, we find null effects for most outcomes that we might expect to mediate the effects of new knowledge about the opposing party. While we hypothesized that both voter and elite corrections would have significant impacts, we actually find diverging effects by party. Among Democrats, only the elite correction reduced predictions of opponents' undemocratic behavior, while only the voter correction was effective among Republicans.[18] Neither correction significantly reduced respondents' sup-

[18] This divergence is particularly surprising given that, in the main study, the treatment event's voter-focused correction reduced Democrats' predictions of norms violations, but not Republicans'. This party difference seemingly flipped in the 10 weeks since the treatment event. There is, of course, an obvious intervening factor: The 2024 election. Effect heterogeneity by party appears very different after the election compared to before.

ports for their own side's violations. This result also opposes our hypothesis that the elite correction would be more effective at reducing those outcomes than the voter correction. The two corrections had similarly weak, conditional effects.

Results were mixed on our outcome of voter vs elite relative blame. In line with our expectations, the voter correction did make Democrats more likely to blame Republican norms violations on Republican *elites* relative to Republican *voters*. However, we do *not* find a significant mirrored effect among Republican respondents, nor did the elite correction change relative blame among respondents of either party.

Although we anticipated heterogeneous treatment effects by party, the mixed results we observe do not neatly line up with theoretical expectations. We have noted before that the treatment conditions are not identical across respondent party because of the different real world situation each party faced. Democrats were somewhat reassured by evidence that Donald Trump would follow the rule of law, while Republicans did not respond to a correction about Joe Biden. Conversely, our correction to Republicans about Democratic voters was more effective than our correction to Democrats about Republican voters.

Neatly interpreting the reasons for these party differences is beyond the scope of this experiment. What we can demonstrate, though, is how misperception corrections about *elites* can have significantly different effects from corrections about *voters*. The strategic and electoral situation of the respondent's party will affect which form of correction, if either, is more effective in reducing predictions of undemocratic behavior. Depolarization practitioners should therefore consider carefully whether misperception corrections should focus on elites or on voters.

This survey experiment is further grim evidence on the potential for misperception corrections to improve commitment to democratic norms. With two corrections for respondents of each party, we still fail to find any significant change in support for undemocratic practices. Overall, the survey experimental results mirror those of the field experiment.

While meta-perceptions corrections may conditionally reduce predictions of an opponent's undemocratic behavior, these corrections have no effect on one's own likelihood of supporting norms violations.

5.2 Depolarization Video Preferences

The final component of the followup survey provided another test of our hypotheses – that Republicans are less interested than Democrats in depolarization initiatives. At the end of the survey, we randomized respondents to see an embedded video of either a) a news clip about a nonpartisan depolarization initiative, or b) a news clip featuring the respondent's co-partisans commenting on the 2024 presidential election results. All videos were of similar length. Respondents were informed that watching the video was optional and would not affect their compensation for the survey. While our field experiment explored compliance gaps for a single treatment, here we test the causal effect of an explicit depolarization video compared a partisan alternative.

Figure 9 shows the difference in video watching duration between respondents given the depolarization video and respondents given the co-partisan video, broken down by respondent party and then by pre-treatment meta-perceptions.[19] We find further evidence that Republicans are less interested than Democrats in depolarization initiatives. Additionally, pooling across both parties, respondents with more hostile pre-treatment meta-perceptions are less interested in depolarization content relative to co-partisan content, a finding that differs from our field experimental results (albeit this survey estimator is exploring the gap between a partisan and depolarization event while the experiment assessed only a single, descriptive compliance estimate). We repeat these tests with a binary outcome

[19] As measured at the beginning of the followup survey.

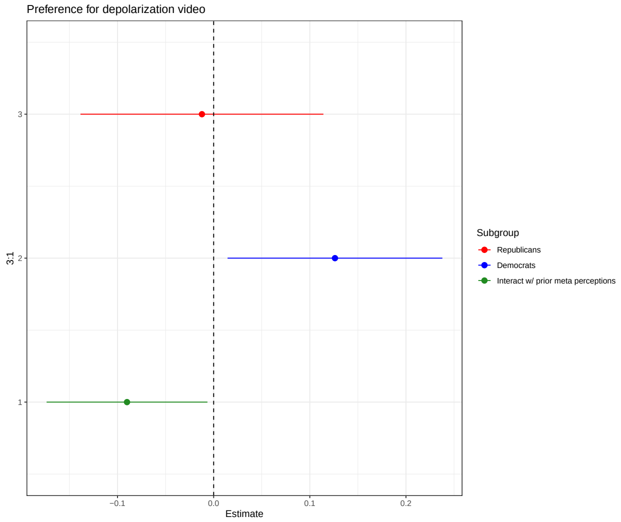

Figure 9: Preference for depolarization video over copartisan video

6 Discussion

In this paper, we assessed a field experimental test of a depolarization event intended to reduce meta-perceptions as a mechanism for improving subject's commitment to democracy and willingness to engage in pro-democracy behaviors. We find that our event modestly reduces meta-perceptions and these results were consistent across party lines and modestly larger for those whose priors were further from the truth. Further, the treatment effect on meta-perceptions demonstrated substantial persistence over time, despite the ephemeral nature of such beliefs (Dias et al., 2024). Despite this promising effect on our primary mediator and a related outcome of predictions of opposing partisan misbehavior, we find no effect on any substantive attitude or pro-democracy behavior of interest, besides increasing reported willingness to attend similar events in the future.

We also report results about attendance and compliance. Despite a generous financial offer to incentivize treatment up-take, we observe a substantial partisan gap in compliance, with Democrats substantially more likely to attend the event. We confirm these results with a follow-up survey experiment on the same sample, showing that Democrats prefer a depolarization event to a co-partisan event, while Republicans were indifferent. For the main event, no pre-treatment attitudinal measure predicted attendance, while individuals with exaggerated meta-perceptions negatively predicted depolarization event attendance in the follow-up experiment. Finally, attending the main treatment event strongly increased willingness to attend future depolarization events, suggesting a combination of generous compensation and successful event experience contributed to future openness to depolarization events.

However, the practical lessons regarding selection into depolarization events are complex. While differential compliance was observed, the magnitude of this gap is modest compared to the overwhelming partisan gaps in depolarization event attendance in the field. This suggests that selection conditional on an offer being made may play a modest role in differential attendance, with differential exposure to the communities and social circles through which depolarization events propagate being a more substantively significant contributor to partisan gaps in depolarization event exposure. Further, the limited evidence of attendance gaps by pre-treatment attitudes suggests that depolarization initiatives may not particularly struggle to attract individuals who hold the attitudes they seek to reduce, though again this stage of selective uptake may be one among many mechanisms contributing to differential exposure.

Our findings suggest that being invited to attend an event, and attending and being compensated for that event, contributes to greater expressed willingness to attend a future event. This mechanism is a costly method to increase attendance, though some forms of financial incentive are common when inducing event attendance and our payments were not unusual given the length of the event. Regardless, the partisan gap that we identify across Republicans' willingness to attend is a deeply important dimension of meta-perception corrections and empirical work generally on partisans. We identify a previously less observable dimension in the external validity of these depolarizing interventions on partisan attitudes.

Finally, we draw on our field and survey experimental results to posit preliminary substantive explanations for why meta-perception corrections did not translate into reductions in anti-democratic attitudes or support for pro-democracy behaviors. First, we show that updating on meta-perceptions contributed more modestly to concrete predictions regarding real-world outcomes, with these prediction effects driven only by Democrats. The strategic logic for why meta-perceptions matter (Braley et al., 2023; Druckman et al., 2023) requires respondents to update their beliefs about real world events, rather than merely their evaluations of opposing partisans at the mass level. In addition, as shown in Appendix Table **??**, predictions are noticeably more stable than meta-perceptions.

We see from our prediction outcomes that movement on these questions was noticeably smaller than movement on meta-perceptions. Predictions were also more stable between waves than meta-perceptions. Even as we moved attitudes surrounding the beliefs of out-partisans, we did not meaningfully shift views about how likely the opposing party was to violate democratic norms, with effects on this outcome isolated to Democrats. Future work should investigate how to shift these plausibly more stubborn beliefs. In our follow-up experiment, we show that voters' concrete predictions about the opposing party's behavior, allocation of blame between opposing party elites and voters and support for anti-democratic behavior are resistant to simple survey treatments that mimic our more intensive and bundled main treatment.

Depolarization initiatives in general and meta-perception corrections in particular offer promise in shifting voter attitudes. However, we have shown that these initiatives face challenges: Republicans are substantially less likely than Democrats to select into depolarization initiatives. Further, changing beliefs, even through time-intensive, paid treatments that cause enduring treatment effects, does not necessarily promote pro-democratic attitudes or cause substantially large updating regarding predictions of concrete actions of the opposing party.

References

Douglas J. Ahler and Gaurav Sood. The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences. *The Journal of Politics*, 80(3):964–981, July 2018. ISSN 0022-3816. doi: 10.1086/697253. URL https://www.journals.uchicago.edu/doi/abs/10.1086/697253. Publisher: The University of Chicago Press.

Kevin K. Banda and Joel Sievert. How Filibuster Rhetoric Informs Perceptions of Politicians. *Legislative Studies Quarterly*, 49(3):673–693, 2024. ISSN 1939-9162. doi: 10.1111/lsq.12445. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lsq.12445. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lsq.12445.

Larry M. Bartels and Nicholas Carnes. House Republicans were rewarded for supporting Donald Trump's 'stop the steal' efforts. *Proceedings of the National Academy of Sciences*, 120(34):e2309072120, August 2023. doi: 10.1073/pnas.2309072120. URL https://www.pnas.org/doi/full/10.1073/pnas.2309072120. Publisher: Proceedings of the National Academy of Sciences.

Alia Braley, Gabriel S. Lenz, Dhaval Adjodah, Hossein Rahnama, and Alex Pentland. Why voters who value democracy participate in democratic backsliding. *Nature Human Behaviour*, 7(8):1282–1293, August 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01594-w. URL https://www.nature.com/articles/s41562-023-01594-w. Publisher: Nature Publishing Group.

David E. Broockman, Joshua L. Kalla, and Sean J. Westwood. Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not. *American Journal of Political Science*, 67(3):808–828, 2023. ISSN 1540-5907. doi: 10.1111/ajps.12719. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12719. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12719.

Jacob R. Brown and Ryan D. Enos. The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*, 5(8):998–1008, August 2021. ISSN 2397-3374. doi: 10.1038/s41562-021-01066-z. URL https://www.nature.com/articles/s41562-021-01066-z. Publisher: Nature Publishing Group.

Leonad Bursztyn and David Yang. Misperceptions About Others | Annual Reviews. 14: 425–452, 2022. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-051520-023322.

Nicholas C Dias, Laurits F Aarslew, Kristian Vrede Skaaning Frederiksen, Yphtach Lelkes, Lea Pradella, and Sean J Westwood. Correcting misperceptions of partisan opponents is not effective at treating democratic ills. *PNAS Nexus*, 3(8):pgae304, August 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae304. URL https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgae304/7730165.

James N. Druckman, Suji Kang, James Chu, Michael N. Stagnaro, Jan G. Voelkel, Joseph S. Mernyk, Sophia L. Pink, Chrystal Redekopp, David G. Rand, and Robb Willer. Correcting misperceptions of out-partisans decreases American legislators' support for undemocratic practices. *Proceedings of the National Academy of Sciences*, 120(23):e2301836120, June 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2301836120. URL https://pnas.org/doi/10.1073/pnas.2301836120.

Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, Linda J Skitka, Joshua A Tucker, Jay J Van Bavel, Cynthia S Wang, and James N Druckman. Political Sectarianism: A Dangerous Cocktail of Othering, Aversion, and Moralization. 2024.

Rebecca Hanson, Dorothy Kronick, and Tara Slough. Preaching to the Choir: A Problem of Participatory Interventions. *The Journal of Politics*, 87(2):739–756, April 2025. ISSN 0022-3816, 1468-2508. doi: 10.1086/732983. URL https://www.journals.uchicago.edu/doi/10.1086/732983.

Gretchen Helmke, Mary Kroeger, and Jack Paine. Democracy by Deterrence: Norms, Constitutions, and Electoral Tilting. *American Journal of Political Science*, 66(2):434–450, 2022. ISSN 1540-5907. doi: 10.1111/ajps.12668. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12668. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12668.

Mohamed A. Hussein and S. Christian Wheeler. Reputational costs of receptiveness: When and why being receptive to opposing political views backfires. *Journal of Experimental Psychology: General*, 153(6):1425–1448, 2024. ISSN 1939-2222. doi: 10.1037/xge0001579. Place: US Publisher: American Psychological Association.

Diana Jordan, Trent Ollerenshaw, and Andrew Trexler. Repeated Measure Designs are Superior for (Most) Experimental Survey Research Applications. 2025.

Matthew Levendusky. *Our Common Bonds: Using What Americans Share to Help Bridge the Partisan Divide*. University of Chicago Press, March 2023. ISBN 978-0-226-82469-7.

Matthew S. Levendusky. Why Do Partisan Media Polarize Viewers? *American Journal of Political Science*, 57(3):611–623, 2013. ISSN 1540-5907. doi: 10.1111/ajps.12008. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12008.

Steven Levitsky and Daniel Ziblatt. *How Democracies Die - Google Books*. Penguin, New York, 2018. URL https://www.google.com/books/edition/How_Democracies_Die/VZKADwAAQBAJ?hl=en&gbpv=1&dq=how+democracies+die&pg=PA1&printsec=frontcover.

Adeline Lo, Jonathan Renshon, and Lotem Bassan-Nygate. A Practical Guide to Dealing with Attrition in Political Science Experiments. *Journal of Experimental Political Science*, 11(2):147–161, 2024. ISSN 2052-2630, 2052-2649. doi: 10.1017/XPS.2023.22. URL https://www.cambridge.org/core/product/identifier/

S2052263023000222/type/journal_article.

Lilliana Mason. *Uncivil Agreement: How Politics Became Our Identity.* University of Chicago Press, April 2018. ISBN 978-0-226-52468-9. Google-Books-ID: R29RDwAAQBAJ.

Joseph S. Mernyk, Sophia L. Pink, James N. Druckman, and Robb Willer. Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, 119(16):e2116851119, April 2022. doi: 10.1073/pnas.2116851119. URL https://www.pnas.org/doi/full/10.1073/pnas.2116851119. Publisher: Proceedings of the National Academy of Sciences.

Samantha L. Moore-Berg and Boaz Hameiri. Improving intergroup relations with meta-perception correction interventions. *Trends in Cognitive Sciences*, 28(3):190–192, March 2024. ISSN 1364-6613. doi: 10.1016/j.tics.2024.01.008. URL https://www.sciencedirect.com/science/article/pii/S1364661324000081.

Lilla V. Orr and Gregory A. Huber. The Policy Basis of Measured Partisan Animosity in the United States. *American Journal of Political Science*, 64(3):569–586, 2020. ISSN 1540-5907. doi: 10.1111/ajps.12498. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12498. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12498.

Lilla V. Orr, Anthony Fowler, and Gregory A. Huber. Is Affective Polarization Driven by Identity, Loyalty, or Substance? *American Journal of Political Science*, 67(4):948–962, 2023. ISSN 1540-5907. doi: 10.1111/ajps.12796. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12796. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12796.

Neil A. O'Brian and Thomas Bradley Kent. Partisanship and Trust in Personal Doctors: Causes and Consequences. *British Journal of Political Science*, 55:e34, 2025. ISSN 0007-1234, 1469-2112. doi: 10.1017/S0007123424000607. URL https://www.cambridge.org/core/product/identifier/S0007123424000607/type/journal_article.

Folco Panizza, Eugen Dimant, Erik O Kimbrough, and Alexander Vostroknutov. Measuring norm pluralism and perceived polarization in US politics. *PNAS Nexus*, 3(10):pgae413, October 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae413. URL https://doi.org/10.1093/pnasnexus/pgae413.

Michael H. Pasek, Lee-Or Ankori-Karlinsky, Alex Levy-Vene, and Samantha L. Moore-Berg. Misperceptions about out-partisans' democratic values may erode democracy. *Scientific Reports*, 12(1):16284, September 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-19616-4. URL https://www.nature.com/articles/s41598-022-19616-4. Publisher: Nature Publishing Group.

Vicente Valentim. Norms of Democracy, Staged Democrats, and Supply of Exclusionary Ideology. *Comparative Political Studies*, page 00104140241283009, October 2024. ISSN 0010-4140. doi: 10.1177/00104140241283009. URL https://doi.org/10.1177/00104140241283009. Publisher: SAGE Publications Inc.

Jan G. Voelkel, Michael N. Stagnaro, James Y. Chu, Sophia L. Pink, Joseph S. Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G. Allen, L. Victor Allis, Gina Baleria, Nathan Ballantyne, Jay J. Van Bavel, Hayley Blunden, Alia Braley, Christopher J. Bryan, Jared B. Celniker, Mina Cikara, Margarett V. Clapper, Katherine Clayton, Hanne Collins, Evan DeFilippis, Macrina Dieffenbach, Kimberly C. Doell, Charles Dorison, Mylien Duong, Peter Felsman, Maya Fiorella, David Francis, Michael Franz, Roman A. Gallardo, Sara Gifford, Daniela Goya-Tocchetto, Kurt Gray, Joe Green, Joshua Greene, Mertcan Güngör, Matthew Hall, Cameron A. Hecht, Ali Javeed, John T. Jost, Aaron C. Kay, Nick R. Kay, Brandyn Keating, John Michael Kelly, James R. G. Kirk, Malka Kopell, Nour Kteily, Emily Kubin, Jeffrey Lees, Gabriel Lenz, Matthew Levendusky, Rebecca Littman, Kara Luo, Aaron Lyles, Ben Lyons, Wayde Marsh, James Martherus, Lauren Alpert Maurer, Caroline Mehl, Julia Minson, Molly Moore, Samantha L. Moore-Berg, Michael H. Pasek, Alex Pentland, Curtis

Puryear, Hossein Rahnama, Steve Rathje, Jay Rosato, Maytal Saar-Tsechansky, Luiza Almeida Santos, Colleen M. Seifert, Azim Shariff, Otto Simonsson, Shiri Spitz Siddiqi, Daniel F. Stone, Palma Strand, Michael Tomz, David S. Yeager, Erez Yoeli, Jamil Zaki, James N. Druckman, David G. Rand, and Robb Willer. Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, 386(6719): eadh4764, October 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adh4764. URL https://www.science.org/doi/10.1126/science.adh4764.

Barry R. Weingast. The Political Foundations of Democracy and the Rule of Law. *The American Political Science Review*, 91(2):245–263, 1997. ISSN 0003-0554. doi: 10.2307/2952354. URL https://www.jstor.org/stable/2952354. Publisher: [American Political Science Association, Cambridge University Press].

Chagai Weiss, Don Green, and Robb Willer. Politicians' Bipartisan Appeals to Civility and Partisan Divides: A Field Experiment with U.S. Governors, May 2025. URL https://osf.io/5qxyw_v1.

Alexander Wuttke and Florian Foos. Making the case for democracy: A field-experiment on democratic persuasion. *European Journal of Political Research*, n/a(n/a), 2024. ISSN 1475-6765. doi: 10.1111/1475-6765.12705. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12705. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6765.12705.

Alexander Wuttke, Florian Sichart, and Florian Foos. Null Effects of Pro-Democracy Speeches by U.S. Republicans in the Aftermath of January 6th. *Journal of Experimental Political Science*, 11(1):27–41, 2024. ISSN 2052-2630, 2052-2649. doi: 10.1017/XPS.2023.17. URL https://www.cambridge.org/core/product/identifier/S2052263023000179/type/journal_article.

Floyd Jiuyun Zhang. Political endorsement by Nature and trust in scientific expertise during COVID-19. *Nature Human Behaviour*, 7(5):696–706, March 2023. ISSN 2397-

3374. doi: 10.1038/s41562-023-01537-5. URL https://www.nature.com/articles/s41562-023-01537-5.

7 Appendices

7.1 Pre-analysis plans and deviations

We filed a pre-analysis plan on OSF[20] before the treatment event. We note two deviations from our pre-analysis plan. First, we changed one of our quasi-behavioral measures. The initial pilot of our outcomes survey revealed that many respondents believed the original quasi-behavioral measure (providing information to write a letter to Congress in favor of funding bridging initiatives) asked for too much personal information - which they mentioned through the feedback mechanism on Cloud Research Connect. We thus replaced this item with a question asking about interest in a future depolarization event. Because of this change, respondents who answered the earliest pilot run of the outcome survey (N=37) do not have a recorded response to the interest in depolarization event outcome, reducing our observed N for that outcome to 3,149. Second, because free speech became a major theme of the unscripted treatment event, we added an additional outcome question battery of the respondent's willingness to defend free speech.

Before our followup survey experiment, we filed an additional pre-analysis plan on OSF. While there were no deviations from this followup pre-analysis plan, we note one ambiguity: We did not clearly specify whether we would be analyzing survey experiment results separately by respondent party ID or pooled across party ID. As question wordings differ meaningfully by the respondent's party, we analyze the results separately by party.

7.2 Hypotheses

For our main experiment, our hypotheses were as follows:

- H1A: Lower compliance rate among Republican respondents (paid sample only)
- H1B: Lower compliance rate among the more affectively polarized

[20] https://osf.io/ausqd , DOI 10.17605/OSF.IO/AUSQD

- H2A: Overall negative treatment effects on support for undemocratic practices and affective polarization, and positive treatment effects on trust in upcoming elections and belief in other party's commitment to democracy (paid sample only)
- H2B: Overall positive treatment effects on our behavioral outcomes (both samples)
- H3A: Higher CACEs for Republicans than Democrats (paid sample only)
- H3B: Higher CACEs for individuals with higher pre-treatments SUPs (paid sample only)
- H3C: Higher CATEs for individuals with meta-perceptions further from the truth

For the follow-up experiment, our hypotheses were:

- H1: Our main treatment effect on meta-perceptions will persist among respondents of both parties
- H2A: Compared to a neutral control, providing information about opposing party elites' commitment to democracy will reduce predictions of norm violations by the opposing party and reduce support for undemocratic practices
- H2B: Compared to a neutral control, providing information about opposing party voters' commitment to democracy will reduce predictions of norm violations by the opposing party and reduce support for undemocratic practices
- H2C: Compared to a neutral control and the polling treatment, information about elites following norms will reduce predictions of norm violations.
- H2D: The polling correction will cause a reduction in the blame voters allocate to out party voters while the elite condition will cause a reduction in the blame voters allocate to out-party elites
- H2E: Respondents with a more pessimistic prior belief about the opposing party's commitment to democracy will have differentially stronger treatment effects from both

the polling and elite conditions, compared to a neutral control.

- H3: Democrats will spend more time than Republicans watching a video about a bridging organization, compared to a comparable-length video where co-partisans discuss the election

Both sets of hypotheses were reported in the respective pre-registrations.

7.3 Descriptive Statistics

Below, we show descriptive statistics for our main sample and the subset that completed the follow-up experiment. We note that the sample is not, and was not designed to be, representative of the American public. However, the low levels of anti-democratic attitudes and a mean 51% affective polarization score are comparable to the national average, per the polarization research lab's polling.[21].

Table 8: Followup Survey Experiment

Statistic	N	Mean	St. Dev.	Min	Max
Republican	2,190	0.418	0.493	0	1
Age	2,190	42.800	13.700	18	83
Nonwhite	2,190	0.244	0.430	0	1
College degree	2,190	0.581	0.494	0	1
Male	2,190	0.419	0.494	0	1
Pre-treatment meta-perceptions	2,187	24.900	27.300	0.000	100.000
Pre-treatment affective polarization	2,190	52.500	30.600	−100	100

7.4 Expanded summary of treatment event

The following is an AI summary of the raw transcript of the treatment event, generated by GPT 4.5.

Throughout the discussion, speakers focused on critical themes such as freedom of

[21] https://americaspoliticalpulse.com/citizens/

speech and the impacts of cancel culture. They emphasized freedom of speech as fundamental to democracy, arguing that open dialogue and the free exchange of diverse viewpoints are essential for maintaining a healthy democratic society. Conversely, they criticized cancel culture as contrary to these principles, stressing that suppressing unpopular or minority views undermines the very foundations of democratic debate.

Another central theme of the evening was political polarization, specifically distinguishing emotional polarization—animosity or distrust towards opposing political groups—from ideological polarization. The speakers acknowledged that many Americans harbor stronger feelings against political opponents than actual disagreements on policy might suggest. To counteract this dynamic, they highlighted the importance of intellectual humility: actively listening, respecting differing views, and thoughtfully engaging with those holding opposing political convictions.

The speakers also shared personal narratives, illustrating their distinct pathways toward political identity. One speaker explained his conservative viewpoints through the experiences of his parents, whose pursuit of the American Dream through hard work, resilience, and personal responsibility profoundly shaped his perspective. In contrast, the other speaker, a pastor, described how his encounters with racial injustice and his advocacy for social justice and equity led him toward liberal perspectives. These personal accounts illustrated how deeply personal experiences can inform political ideologies.

Despite their ideological differences, both speakers consistently underscored mutual respect, friendship, and a commitment to finding common ground. They agreed that most Americans genuinely seek peaceful interactions and constructive political dialogue, rather than conflict or hostility. Audience engagement was actively encouraged during the event, with attendees invited to submit questions to deepen the exploration of the issues discussed. Overall, the event fostered civility, mutual understanding, and the shared objective of strengthening democratic discourse.

7.5 Full Models

7.5.1 Main outcomes, ITT

7.5.2 Main outcomes, CACE

Table 9: CACE estimates

	Metas	Free speech	Behaviors	Undemocratic	Aff Pol.	Predictions
Treatment	−0.199***	−0.007	0.026	−0.011	−0.256	−3.267*
	(0.048)	(0.054)	(0.017)	(0.041)	(0.878)	(1.398)
Num.Obs.	3180	3180	3155	3180	3180	3180
R2	0.274	0.113	0.023	0.072	0.753	0.255
R2 Adj.	0.270	0.109	0.018	0.068	0.752	0.251

$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 10: CACE on meta-perceptions estimates by differing compliance definitions

	(1)	(2)	(3)
Watched	−0.199***		
	(0.048)		
Watched and answered correctly		−0.210***	
		(0.051)	
Attended event live only			−0.422***
			(0.103)
Num.Obs.	3180	3180	3180
R2	0.274	0.274	0.262
R2 Adj.	0.270	0.270	0.259

$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

7.5.3 Difference-in-means models

7.5.4 Followup survey outcomes

[htbp!]

Table 11: Followup Survey Experiment, Democratic Sample

	Dem. Predictions of Violations	Opponents' fairness	Support for violations	Blames leaders (dif)
Voter correction	−0.047	0.056	−0.108	0.165**
	(0.061)	(0.062)	(0.094)	(0.055)
Elite correction	−0.212***	−0.020	−0.120	0.031
	(0.061)	(0.062)	(0.094)	(0.055)
Num.Obs.	1269	1269	1269	1269
R2	0.188	0.158	0.119	0.012

$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 12: Followup Survey Experiment, Republican Sample

	Predictions of Violations	Opponents' fairness	Support for violations	Blames leaders (dif)
Voter correction	−0.174*	0.002	−0.039	0.036
	(0.073)	(0.072)	(0.096)	(0.073)
Elite correction	−0.039	0.043	0.094	0.103
	(0.073)	(0.071)	(0.095)	(0.072)
Num.Obs.	912	912	912	912
R2	0.238	0.289	0.278	0.009

$^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 13: Preference for depolarization video

	Republicans	Democrats	Party interaction	Pretreatmen
Bridging video	−0.012	0.126*	0.126*	0.0
	(0.064)	(0.057)	(0.056)	(0.0
Republican			0.082	
			(0.063)	
Bridging video*Republican			−0.138	
			(0.086)	
Pretreatment metaperceptions				0.05
				(0.0
Bridging * Pretreatment meta perceptions				−0.0
				(0.0
Num.Obs.	912	1269	2181	218
R2	0.026	0.030	0.025	0.0
Std.Errors	IID	IID	IID	II

+ p \num{¡ 0.1}, * p \num{¡ 0.05}, ** p \num{¡ 0.01}, *** p \num{¡ 0.001}

Models include demographic covariates

7.5.5 Other alternative specifications

Table 14: Preference for depolarization video, binary outcome

	Republicans	Democrats	Party interaction
Bridging video	−0.106***	−0.021	−0.022
	(0.032)	(0.026)	(0.027)
Republican			0.061*
			(0.030)
Bridging video*Republican			−0.085*
			(0.041)
Pretreatment metaperceptions			
Bridging * Pretreatment meta perceptions			
Num.Obs.	912	1269	2181
R2	0.044	0.047	0.044
Std.Errors	IID	IID	IID

The outcome for these models is a binary indicator of whether a respondent stayed on the video page fo

+ p \num{¡ 0.1}, * p \num{¡ 0.05}, ** p \num{¡ 0.01}, *** p \num{¡ 0.001}

Models include demographic covariates

7.6 Additional Heterogeneous Effect Analyses

7.7 Heterogeneous effects, followup survey experiment

Figure 12: Heterogeneous effects by meta-perceptions as measured at the beginning of the followup survey

Table 15: Party compliance gaps by Block

	Central	Eastern	Florida	Mountain	Pacific/Alaska/Hawaii
Republican	−0.084**	−0.072**	−0.055	−0.064	−0.036
	(0.030)	(0.024)	(0.050)	(0.063)	(0.042)
Num.Obs.	924	1454	342	216	525
R2	0.009	0.006	0.004	0.005	0.001
Std.Errors	IID	IID	IID	IID	IID

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include demographic covariates

Table 16: Heterogeneity for Video Interest by Pre-Treatment Attitudes

	Duration	Duration
Bridging Video	0.068	0.107
	(0.043)	(0.084)
Aff. Pol	−0.001	−0.001
	(0.001)	(0.001)
Meta-Perceptions	0.051+	
	(0.029)	
Bridging:Affpol		−0.001
		(0.001)
Bridging:Metas	−0.091*	
	(0.040)	
Num.Obs.	2184	2181
R2	0.026	0.024

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models includes demographic covariates

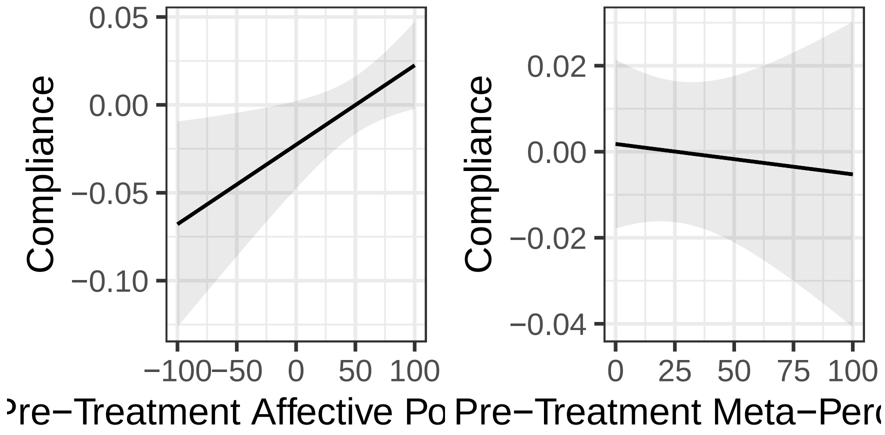

Figure 10: Continuous Pre-Treatment Affective Polarization and Meta-Perception Measures Do Not Strongly Predict Differential Compliance

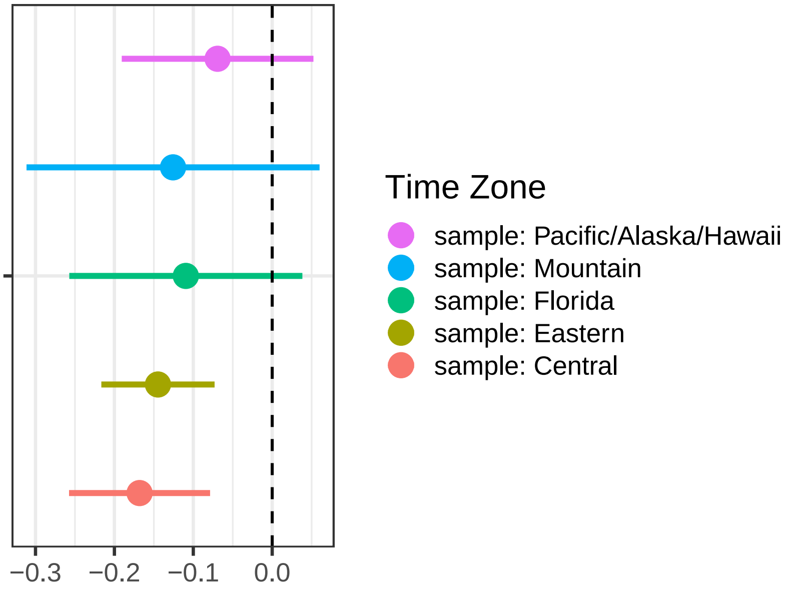

Figure 11: Partisan Attendance Gap by Region

7.8 Mass email invitation experiment

An additional pre-registered arm of this project randomized sending an invitation to the event (but no monetary offer) to half of a 17,806 member email list of a partner organization. Following the event, the full email list received an outcome collection email that described three behavioral outcomes from the main study (the two pledges and poll worker volunteering) and provided links to sign up for them. We track clicks on those links as our sole outcome in this experiment.

Due to a delay from our partner organization, the email containing the outcome links was delayed until two weeks after the treatment event. Perhaps because of this, attention to the emails was unexpectedly low in both treatment and control groups. For our outcome email, We recorded 23 link clicks from the treatment group and 12 in control, a rate of one click per 508 recipients. Although this is borderline statistically significant (χ^2 p value=0.09), the effect size is minuscule. We also cannot confirm whether email receipt lead to event attendance.

In summary, this randomized messaging did not produce meaningful results due to the poor reachability of the sample. We report it anyway in line with our preregistration.

7.9 Attitude Stability Across Measures

Here we explore how pre-treatment beliefs about and attitudes towards the opposing party predict post-treatment outcomes in the control group. Consistent with prior research (Dias et al., 2024), we find in Table 17 that meta-perceptions are relatively weakly held and unstable belief, though we note that predictions about the actions of the opposing party were notably more stable and these beliefs also shifted in response to our treatment (a mechanism which we explored in more detail in our follow-up experiment). As a benchmark, both beliefs were less well predicted by the earlier survey measure than affective polarization, though the predictions were closer in stability to affect than to meta-perceptions. These results reinforce results by Dias et al. (2024) that meta-perceptions show substantial instability, but also suggest that the movement along our prediction outcome might reflect a more durable shift.

Table 17: Attitude Stability in Control Group

	(Meta-Perceptions Post)	(Predictions Post)	(Aff-Pol Post)
Meta-Perceptions Pre	0.519***		
	(0.022)		
Predictions Pre		0.742***	
		(0.017)	
Aff. Pol. Pre			0.855***
			(0.012)
Num.Obs.	1589	1592	1592
R2	0.257	0.546	0.748
Std.Errors	IID	IID	IID

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Models include demographic covariates

7.10 Follow-Up Experimental Materials