Changing Beliefs about Opponents as a Field Test of Depolarization Initiatives *

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

January 24, 2026

Abstract

Correcting false and negative beliefs about political opponents has shown promise as a technique for reducing polarization and anti-democratic attitudes. This style of intervention depends on uptake among the populations most likely to display harmful attitudes. To test these mechanisms, we worked with a partner organization 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. Second, our intervention achieved a durable reduction in beliefs that opponents were opposed to democracy. However, we found no reduction in anti-democratic attitudes across many pre-registered outcomes.

Word count: 9769

*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.The pre-analysis plan for this study is available here.

[†]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 working to depolarize Americans have developed promising methods for reducing partisan animosity and correcting harmful misperceptions about opposing partisans. For these techniques to shift public opinion at scale, however, they must reach individuals across party lines who hold dangerous, inaccurate, or hostile beliefs about political opponents and deliver messages that are both generalizable and cost-effective with enduring effects. Achieving this is difficult because the same forces that generate misperceptions in the first place impede outreach: geographic segregation (Brown and Enos, 2021), social sorting (Mason, 2018), heavily partisan mass and online media environments (Levendusky, 2013), and a growing partisan divide in institutional trust that has led many Republicans to view academic and nonprofit organizations with hostility (Zhang, 2023).

One solution to this problem is messages propagated by partisan elites that generate their own earned media. At the same time, elite-driven depolarization efforts often adopt an idealistic, anti-partisan tone that can harm candidates' prospects in primary campaigns or call into question their partisan bona fides. These challenges are especially evident in the case of meta-perception corrections, which involve delivering conciliatory information that opponents are less threatening than they appear. Yet partisan politicians with career concerns may be reluctant to provide conciliatory statements about their opponents, and polarized partisans may be unlikely to opt into settings where they might be exposed to such messages.

To assess the impact of an intensive, elite-led meta-perception correction, we conducted a field experiment in partnership with a Florida-based bridging organization. The intervention bundled a direct meta-perception correction with an hour-long, elite-facilitated demonstration of civil discourse, modeling how actors can affirm a shared commitment to democracy while sustaining sharp disagreement on substantively important issues; participants were compensated for 30 minutes of required attendance. Throughout the event,

speakers repeatedly invoked democratic norms of free speech and the importance of expressing political beliefs without fear of retaliation. The design aimed to approximate a realistic form of political discussion that elites could plausibly adopt without provoking backlash from co-partisans. We assess the intervention's effects on meta-perceptions of out-partisans and support for anti-democratic actions, especially regarding free speech and censorship, and include quasi-behavioral measures capturing respondents' willingness to engage in pro-democratic behavior beyond the survey.

We draw from work on depolarization initiatives and meta-perceptions corrections (Hanson et al., 2025). Qualitative accounts from practitioners and prior academic research identify three challenges to the practical effectiveness of depolarization initiatives: 1) The ephemeral nature of democratic meta-perceptions; 2) The difficulty in creating depolarization messaging that self-interested partisan politicians are willing to repeat at scale; and 3) Self-selection into depolarization events, rendering them difficult to administer interventions to the subjects who hold hostile attitudes at baseline. We design our treatment to investigate all three of these pitfalls. Our intervention consists of a memorable, 30-minute long virtual town hall-style event, with real political elite messengers who demonstrate respect for democratic norms amid sharp issue disagreement. By using an online sample of compensated participants, we use a sample that is unlikely to select into depolarization events, similar to other efforts to test bridging initiatives in the field, for example by paid advertisements in streaming services (Weiss et al., 2025).

To investigate the practical challenges of scaling depolarizing interventions, our design allows for a pre-registered assessment of compliance among participant sub-groups. We do this to explore 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 experi-

ment to derive an additional causal estimate of differential compliance with depolarization treatments.

Ultimately, our bundled treatment successfully and enduringly reduces meta-perceptions that opposing partisans supported antidemocratic behaviors. These intent-to-treat effects are substantively similar across subgroups, despite a partisan gap 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: those holding the most harmful attitudes do not appear to be systematically selecting out of treatment. Intentions to participate in future depolarization events increase with event-offers and attendance. We conclude that depolarization events can work to durably reduce negative beliefs about out-partisans, while also increasing the likelihood of future event attendance. However, we find no clear evidence that such events change first order attitudes.

2 Background

2.1 Democratic Meta-Perceptions

Substantively, our experiment is rooted in a literature on second-order beliefs about opposing partisan's commitment to democracy, known 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

[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)

similar theoretical axes.

Support for democratic backsliding at elite and mass levels is 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 opponents' retaliation can constrain backsliding (Weingast, 1997; Miller, 2021; Helmke et al., 2022). 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 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) in a mass survey 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

[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)

majority of experimental subjects.

The first obstacle to enduringly changing attitudes is rooted in the nature of meta-perceptions themselves, as highlighted in a recent study (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 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 (Valen-

tim, 2024). However, 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, to improve ecological validity, we test 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 (Dias et al. (2024)), messages delivered by politicians of their own accord are easier to scale, especially because prominent figures will generate earned media coverage. Further, these messages constitute a special case of elite opinion leadership directed at promoting pro-democratic attitudes (Wuttke and Foos, 2024; Wuttke et al., 2024).

Finding suitable elite messengers is a challenge, however, 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 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 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 across the U.S., we deliver our treatment event as part of a real political discussion that did not ignore strident partisan disagreement on issues. We did not require elite actors to repudiate their party nor call their partisan credentials into dispute.

2.3 Selection into Depolarization

Most significantly, we consider the problem of selection into receiving depolarization messaging. From our background interviews, the organizers of public depolarization 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." Although 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 records of event participants at 19 depolarization organizations (total N of 1,915 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 survey experiments where non-compliance is rare Voelkel et al. (2024) or experiments that condition on explicit willingness to attend a specific event.

Conceptually closest, albeit in a dramatically different context, is a recent paper that assessed whether town halls linked to community policing efforts risked "preaching to the choir" insofar as attendees were not in need of reassurance about the propriety of police behaviorHanson 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).

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 factions, even Trump-supporting Republicans normally unreachable by depolarization initiatives. 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 of behavioral evidence of actual event attendance.

3 Experimental Design

To test both the effectiveness of our 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 virtual 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=2,181 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 an unspecified 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 randomized by block into treatment (N = 1,730) and control (N = 1,731) by party * region block. The 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)[3]. 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. Given that the live event was offered at a specific time, the 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. Table 9 in the Appendix shows balance between assignment groups.

[3]Florida is given its own block as our partner organization and speakers are based there.

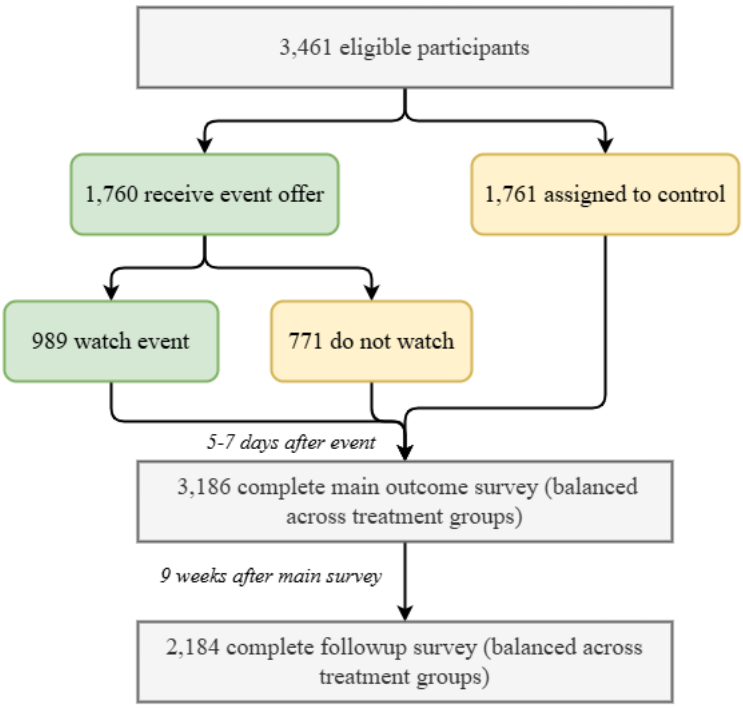


Figure 1: Experimental design

3.2 Treatment Event

One week after the screening survey, we invited the treatment group to attend the treatment event on Thursday, September 18th, 2024 at 7 PM EDT. 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. 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. We closed access to the recorded event 48 hours after the live broadcast.

The treatment consisted of a public conversation between a Republican state rep-

[4]This price was very generous compared to standard Cloud Rates which are usually smaller in both hourly rate and total compensation

resentative and a Democratic local religious leader that blended free-flowing dialogue with explicit reminders that support for democratic norms is bipartisan. Early in the event, participants were asked to guess answers to our meta-perception items, after which we provided the correct figures from existing polling and our own pre-treatment survey. These factual reminders, which approximate a survey ask–tell intervention, initiated a broader discussion about respect and listening across party lines. The conversation foregrounded the speakers' life stories and friendship despite partisan differences, as well as their shared commitment to American democracy and to defending the free speech of political opponents. After 30 minutes, the conversation shifted to specific policy issues such as immigration. At that point, respondents were free to leave the event and receive full payment. Respondents appear to have been attentive and interested: 89% of attendees correctly answered both attention-check questions, and 15% continued to watch the event for at least 15 minutes longer than required. We also received numerous direct messages from attendees reporting that they had genuinely enjoyed the event. In the typology of Moore-Berg and Hameiri (2024), our treatment was a "framed" intervention, providing factual information embedded in a narrative of bipartisan commitment to democracy and combining both facts and parasocial contact. In modeling a civil discussion in this parasocial atmosphere, our treatment event could be compared to group deliberation interventions that have achieved enduring treatment effects (Mendelberg, 2002; Fishkin et al., 2024).

3.3 Outcome surveys

We launched the main outcome survey 8 days after the event, keeping it open for the next 3 days; an additional opportunity (sent via reminder to respondents who had not completed the survey) 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.[5] We do

[5]Given that respondents were allowed to watch the recorded treatment event up to 48 hours after the live event, respondents completed the outcome survey between 6 and 15 days after receiving treatment.

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 connection with the treatment event or the pre-treatment screener survey (conducted about two weeks before the event)[6]. In addition to the temporal 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) suggest that there is little risk of consistency bias. Instead, we gain considerable statistical power from highly prognostic pre-treatment covariates. We achieved a recontact rate of 92% in the control group and 93.5% in the treatment group, a gap that is not statistically significant[7].

3.4 Outcome measures

We measured a wide range of survey and quasi-behavioral outcomes to test for changes in both attitudes and real-world behaviors. We list our outcome measures of interest in Table 1. The full survey instrument is available in Appendix Section 7.1 and is summarized in 1. We chose our questions on undemocratic practices from the broad set of plausible options (Ahmed, 2022) in order to limit ceiling effects (where all respondents disapprove of a practice in the control group) and be conceptually relevant to our treatment event. Despite this, 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. To move beyond purely attitudinal measures, we include four quasi-behavioral outcomes. Two ask respondents to provide their name and state for inclusion in public pledges, and two ask for permission to share contact information with an outside organization.[8]

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

[7]Re-contact procedures are discussed in the appendix

[8]We conservatively code responses with obviously false names or contact information as zero.

Outcome	Questions
Meta-perceptions battery	Average of four questions asking what percentage of opposite party members agree with the following:[9] • 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	Same questions as meta-perceptions battery, with own party and opposite party reversed
Affective polarization	Gap in feeling thermometer ratings of own vs opposing 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.*[10] • 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 1: Primary outcome questions

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, while 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 pre-analysis plan. We first discuss our 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_i is a vector of preregistered covariates (education, race, partisanship, presidential vote choice, and pretreatment attitudes), and τ_i is a vector timezone-party dummy variables.

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

4.1 Compliance

We begin with pre-registered compliance results. The first-stage outcome is observed without attrition because failing to comply with treatment assignment 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. [11]

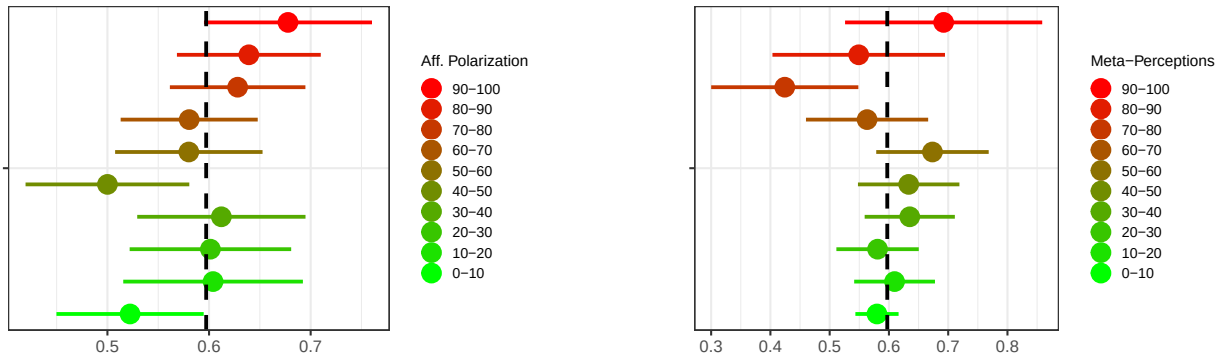
[11] Compliance can be defined as watching 30 minutes (our preregistered measure), watching and passing both attention checks (55.7%), or watching live only (27.8%), and the main compliance-rate patterns are substantively unchanged across these definitions.

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 8).

Next, we explore how compliance is predicted by attitudinal pre-treatment covariates. 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 participation is predicted by these attitudinal measures. In Figure 2, we show that there is little evidence of a strong relationship between 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 and 7, though this difference does not approach statistical significance across any model.

Because compliance is observed among a sample who regularly engage in paid tasks, we interpret non-compliance as particular aversion to a well-compensated event; the treatment event paid substantially better on a per-hour and overall basis than the majority of

[12]To preserve the ecological validity and verisimilitude of the treatment event, our partner organization's website described 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.



(a) Compliance by Affective Polarization

(b) Compliance by Meta-Perceptions

Figure 2: Heterogeneous Effects of Compliance. The figure displays compliance rates conditioned on affective polarization (a) and meta-perceptions of opponents (b).

tasks on the platform. These first stage results are robust to alternatively defining compliance as attending and correctly answering both attention checks. Throughout, we treat the experimental design as a case of one-sided non-compliance rather than as an encouragement design because while the event was advertised outside of incentivized recruitment, regular online attendance is minimal (between 50-100 people) and unlikely to overlap demographically with the experimental universe.

4.2 Main Results

We next estimate our main results, using both intent-to-treat (ITT) and complier-average-causal-effect (CACE) estimates. We calculated CACE estimates in R with a two-stage least squares model using the ivrobust function in the estimatr package. While there are several possible definitions of compliance, these produce small changes to the estimate and our results are robust to these alternative specifications. The different specifications of compliance did not change which effects are statistically significant, nor did they cause substantively large changes to CACE estimates.

We first report our four main pre-registered survey outcomes. Because of the specific structure of the intervention, the meaning of the ITT is the effect of random assignment

to 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 a dummy variable for timezone-party blocks.

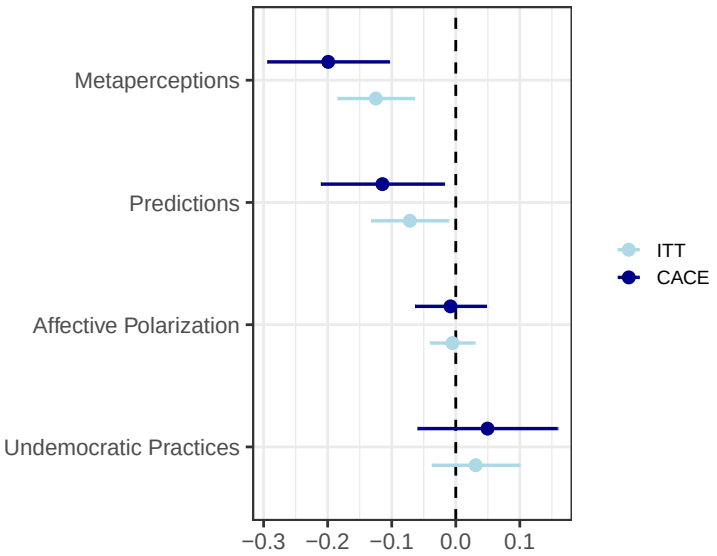


Figure 3: Treatment Changes Meta-Perceptions but not Attitudes

We begin by testing whether the event's meta-perception correction affected second-order beliefs at the time of our initial outcome survey, with this meta-perception battery serving as our primary mechanism check. This is a more difficult test of belief change than many survey experimental tests of meta perception correction. First, the outcome survey followed the event by 5 to 10 days, far longer than the near-instantaneous sequence of treatment and outcome questions in many survey experiments. Second, while many "ask-tell" style corrections have features that may induce demand effects above and beyond the standard risks of such effects in surveys, including congratulatory messages for correct guesses or comparison tables that highlight the gaps between guesses and true estimates. To mitigate these effects our outcome survey was not overtly linked to the treatment event, was

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

not advertised prior to or during the event, and was administered by a researcher account with a different name.

Figure 3 shows modest but significant ITT effects on meta-perceptions (0.12 SD), with larger though more imprecisely estimated effects for compliers (.199 SD). Treated participants learned from, and retained, the correction for a longer period than standard survey experiments are able to assess. Effects appear on all four meta-perception items, with the largest for "[Opposite party] should use violence against peaceful protests by [Own party]". We did not pre-specify item-level hypotheses. The treatment also reduced predictions of post-election anti-democratic behavior ($p < 0.05$), making respondents more optimistic that the opposing party *would not* violate norms; we explore significant heterogeneity in this effect later. By contrast, we find *no treatment effects* on broader political attitudes: support for own-party undemocratic practices (censorship or unfair electoral tactics), support for the out-party's free-speech rights, or affective polarization. These nulls remain before multiple-comparison adjustments, with small point estimates, and hold across two additional outcome sets reported in the appendix. The contrast is striking given the survey design: attitude batteries followed immediately after (though on a separate page from) the meta-perception items, which should have encouraged reciprocal reasoning; one null battery mirrored the treated meta-perception battery with only party labels swapped, the kind of proximate outcome prior work suggests should move (Dias et al., 2024). As noted in the pre-analysis plan, the study was well powered to detect effects as small as $\frac{1}{10}$ SD.

Did our treatment and its effects on meta-perceptions translate into behavioral outcomes? In Figure 4, we show ITT effects for our four quasi-behavioral measures and a combined index of the four. The treatment had a null effect on the combined index, but with considerable variation between individual behaviors. We find a large and significant effect on respondents' interest in attending a future depolarization event, but no effect on pro-democratic behaviors.

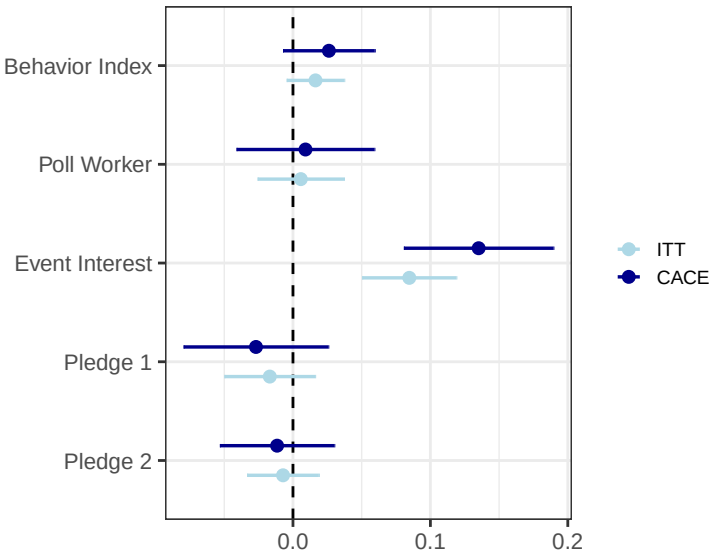


Figure 4: Treatment Changes Interest in Future Events but no Other Behavioral Outcomes

4.3 Heterogeneous Effects

We pre-registered three predictions of heterogeneous treatment 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.

Next, we test for pre-registered heterogeneous effects in our estimates across two pre-treatment attitudinal variables: partisanship and pre-treatment attitudinal measures. Full models are available in the appendix, but we find no evidence of treatment effect heterogeneity across any outcome. The sole evidence of heterogeneity is a modest interaction between pre-treatment meta-perceptions and the effect on the prediction outcomes, such that participants with higher meta-perceptions were less re-assured by treatment about the intentions of the opposing party.

Our treatment event provided respondents with correct statistics about opposite party beliefs, bundled with messages relating to civility and misperceptions in general. If

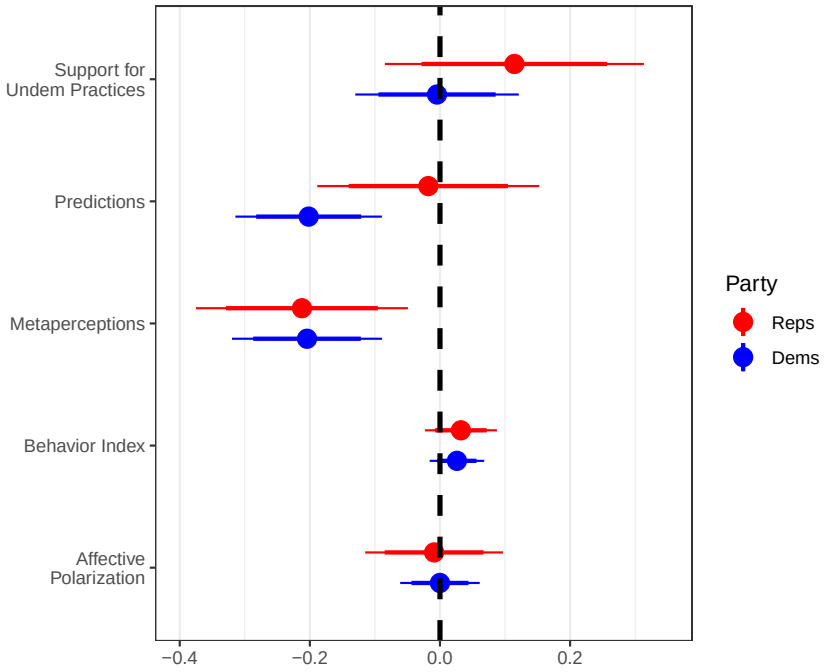


Figure 5: CACEs by respondent party for main outcomes

Table 2: 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 3: 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

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 Table 2, 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.

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 effect 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%.

4.4 Attrition

Our design faced a risk of attrition common to similar experimental designs (Lo et al., 2024). Further, practical considerations that rendered a placebo infeasible,[14] 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 a generous offer of compensation to attend our treatment event (specifically, the risk was that receiving the payment would cause some respondents to be if-treated reporters, because they would otherwise have stopped taking surveys on Connect, but were motivated to remain on the platform due to the generous compensation. 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. The intuition behind potential treatment-induced differential attrition would be 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.

Ultimately, only 8% of participants failed to complete the outcome survey and treatment assignment did not predict completion of the outcome survey. Treatment compliers exhibited lower attrition than non-compliers in the treatment group, but attrition was balanced overall across treatment and control conditions. Our reporting results suggest that online recruitment platforms paired with generous incentives can achieve low and balanced attrition, as offers are sufficiently attractive to limit attrition among most respondents and there is minimal (though non-zero) churn among the participant pool over short periods.

Attrition is modestly predicted by partisanship, with Republicans being 6% less

[14]Notably, the partner organization did not wish to implement a placebo. In addition, we anticipated a high compliance rate which meant a placebo was unlikely to improve statistical power (Gerber and Green, 2012)

Table 4: 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

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 experiments

Our field experiment suggests that more flexible interventions than standard survey-based ask–tell corrections can durably shift meta-perceptions and change predictions about the behavior of the opposing party, but may have limited effects on democratic attitudes and behavior. Further, we found differential compliance by partisanship. We conducted a follow-up experiment in early December 2024, roughly 10 weeks after treatment and three weeks after the 2024 presidential election. Our goal was to investigate both persistence of shifting meta-perception and the mechanisms which prevented these changing beliefs from translating into pro-democratic preferences. We invited the same CloudResearch Connect participants who completed the experimental screener and, consistent with the main outcome survey, provided no overt link to the original treatment. After re-measuring meta-perceptions (described above), the follow-up included an embedded, placebo-controlled experiment correcting misperceptions about undemocratic behavior by either opposite-party voters or opposite-party leaders.

5.1 Persistence of Meta-Perceptions Correction

How long did the treatment event's effect on meta-perceptions endure? While our initial survey was already less immediate than a survey experiment, we also tested for longer-term persistence. Our follow-up survey, conducted two months after the treatment event, began with a re-measurement of respondent meta-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. Table 5 shows the intent-to-treat effect of assignment to the treatment group on the meta-perceptions battery collected during the followup survey.

Table 5: Persistence Results

	Followup survey	Main survey
	−0.064+	−0.125***
	(0.038)	(0.030)
Num.Obs.	2187	3180
R2	0.213	0.272

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

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 after two months, suggesting durable updating during a period of saturated media coverage. We find no evidence of a difference in persistence by party ID. The followup survey also showed null effects of initial treatment on a new pair of meta-perceptions questions about the opposite party's support for undemocratic practices in the aftermath of the 2024 election. 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.

5.2 Follow-Up Survey experiment: Correcting misperceptions about voters vs. about elites

Then, to investigate why shifting meta-perceptions might effect citizen preferences, we randomly assigned respondents to one of three conditions: (1) A control (2) A voter correction condition provided respondents with recent polling data showing that opposite party voters broadly support the rule of law and accept election results.[15] (3) An elite correction condition provided respondents with sourced evidence suggesting that the opposite party's leadership would respect the election results and the rule of the law. Table 27 includes the wordings of treatments and outcome wordings shown to respondents of each party. 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.

After the informational treatment, respondents answered four groups of outcome questions (wording varied by party; see Table 27). Our outcome measures are: (1) an index of three items on expected future undemocratic behavior by the opposing party, (2) relative blame for the opposing party's unethical behavior (leaders vs. voters; higher values indicate more blame on leaders), (3) agreement that the opposing party contested the 2024 election fairly, and (4) support for the respondent's own party taking a norm-violating action (Democrats: attempting to block certification of the 2024 results; Republicans: prosecuting former Biden administration officials once Trump takes office), for which baseline support is substantial (59% among Republicans; 28% among Democrats), mitigating concerns about floor effects.

Figure 6 shows the treatment effects of both corrections (relative to the no-information

[15]This is 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

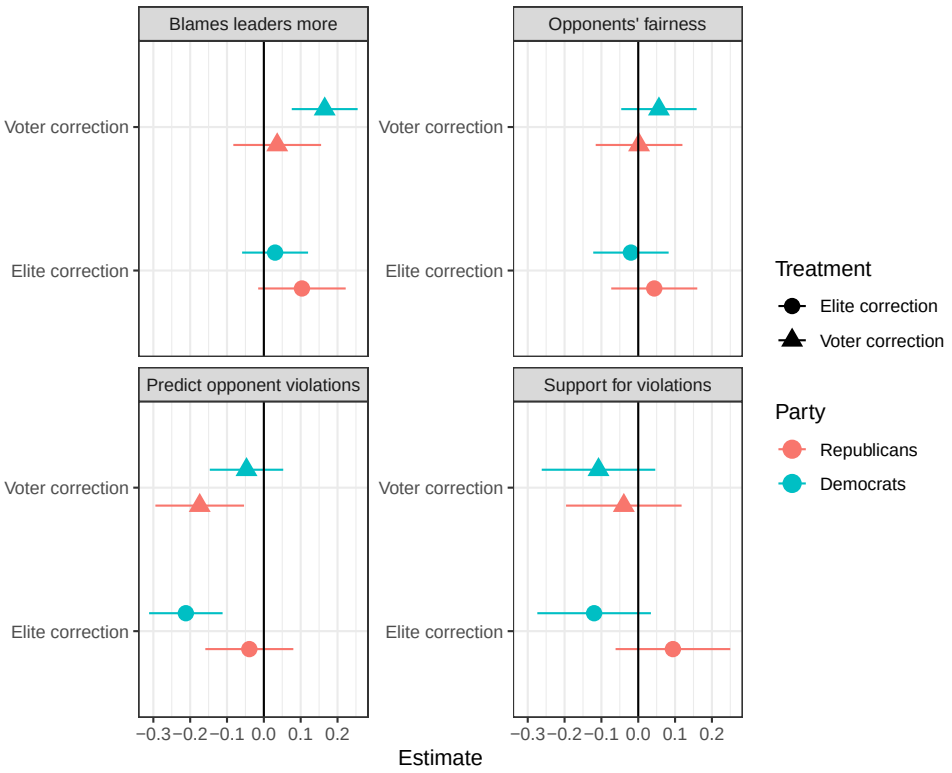


Figure 6: Effects of Elite and Voter Corrections on Main Outcomes

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.[16] Neither correction significantly reduced respondents' supports for their own side's violations. This result also opposes our pre-registered hypothesis that the elite correction would be more effective at reducing those outcomes than the voter

[16]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. Prediction heterogeneity by party appears very different after the election compared to before.

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. 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.

5.3 Depolarization Video Preferences

The final component of the followup survey provided another test of whether Republicans are less interested than Democrats in depolarization initiatives. 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.

Table 6: Democrats Prefer Depolarization, While Republicans are Indifferent

	Republicans	Democrats
Bridging Video	−0.012	0.126*
	(0.065)	(0.057)
Num.Obs.	912	1269
R2	0.026	0.030
Std.Errors	Robust	Robust

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

Table 6 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.[17] 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.

6 Discussion

Through a field experimental test of a depolarization event, we find that our event significantly reduced meta-perceptions that opponents were opposed to democracy. 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 usually ephemeral nature of such beliefs (Dias et al., 2024). Notwithstanding this promising effect on our primary mediator, as well as a related outcome of predictions of opposing partisans' misbehavior, we find no effect on any substantive pro-democracy attitude or behavior of interest, aside from increasing reported willingness to attend similar depolarization events in the future.

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

We also report results about attendance and compliance. Despite a generous financial offer to incentivize treatment uptake, we observe a substantial partisan gap in compliance. Democrats are substantially more likely to attend the event, both overall and across time zone groups. 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 significantly predicted attendance, while exaggerated meta-perceptions of opponents 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 that generous compensation and a positive event experience contributed to future openness to depolarization events. While this result may suggest that positive experiences at depolarization events are habit-forming, similar to other political behaviors (Coppock and Green, 2016), it may also imply that participants are learning about the financial benefits of attending such events.

The practical implications of our findings 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.[18]

[18] A back-of-the-envelope calculation from our compliance result suggests depolarization attendance would be 54% Democratic if offers were made to an equal number of Democrats and Republicans, while the closest

One promising finding regarding attendance is that paying respondents to attend a depolarization event increases expressed willingness to attend similar events in the future. This mechanism is a costly method to increase attendance, although as discussed in our introduction, some forms of financial incentive are common among bridging initiatives. Regardless, the partisan gap that we identify in Republicans' willingness to attend is a deeply important dimension of meta-perception corrections and depolarization initiatives more broadly. The vast majority of studies in this space assess the effect of treatment within a survey experiment or in a context where non-compliance is rare. In fact, many placebo-controlled designs are premised on subjects not knowing the content of an intervention when they choose to participate, removing the possibility of non-compliance being driven by aversion to depolarizing efforts.

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 in our results or in similar studies (Dias et al., 2024). First, we show that updating on meta-perceptions contributed more modestly to concrete predictions about 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. The effectiveness of meta-perception corrections may also depend on whether they concern opposing party elites or voters, and our investigations of these mechanisms in our follow-up experiment were in-conclusive. Notably, our prediction outcomes were also more stable between waves than meta-perceptions, as shown in Appendix Section 7.9. Even as treatment shifted second-order beliefs about out-partisans, it did not shift all

analogy from observational data shows that the ratio of liberals to conservatives in our observational data is 6-to-1. As such, differential uptake on the scale we observed cannot explain more than about 15% of the observed ideological attendance gap.

subgroups' beliefs about real-world outcomes. Future work should investigate how to shift these plausibly more stubborn beliefs.

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.

Amel Ahmed. Is the American Public Really Turning Away from Democracy? Backsliding and the Conceptual Challenges of Understanding Public Attitudes. *Perspectives on Politics*, pages 1–12, July 2022. ISSN 1537-5927, 1541-0986. doi: 10.1017/S1537592722001062. URL https://www.cambridge.org/core/product/identifier/S1537592722001062/type/journal_article.

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.

Alexander Coppock and Donald P. Green. Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities. *American Journal of Political Science*, 60(4):1044–1062, 2016. ISSN 1540-5907. doi: 10.1111/ajps.12210. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12210. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12210.

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.

James Fishkin, Valentin Bolotnyy, Joshua Lerner, Alice Siu, and Norman Bradburn. Can Deliberation Have Lasting Effects? *American Political Science Review*, 118(4):2000–2020, November 2024. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055423001363. URL https://www.cambridge.org/core/product/identifier/S0003055423001363/type/journal_article.

Alan S. Gerber and Donald P. Green. *Field experiments : design, analysis, and interpretation*. W.W. Norton, 2012. URL https://cir.nii.ac.jp/crid/1971149384742718888.

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.

Tali Mendelberg. THE DELIBERATIVE CITIZEN: THEORY AND EVIDENCE. 2002.

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.

Michael K. Miller. A Republic, If You Can Keep It: Breakdown and Erosion in Modern Democracies. *The Journal of Politics*, 83(1):198–213, January 2021. ISSN 0022-3816. doi: 10.1086/709146. URL https://www.journals.uchicago.edu/doi/full/10.1086/709146. Publisher: The University of Chicago Press.

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.

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, May 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01537-5. URL https://www.nature.com/articles/s41562-023-01537-5.

7 Appendix

7.1 Pre-analysis plans and deviations

We filed a pre-analysis plan on OSF 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.

7.1.1 Power analyses

The power analyses in our pre-registrations estimated that our study had nearly 100% power to detect a main effect of 0.2 Cohen's d on the combined 4-item scale and over 80% power to detect an interaction effect of 0.2 Cohen's d. The full power analysis is available in the linked pre-registration.

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
- 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 (Table 7) and the subset that completed the follow-up experiment (Table 8). 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.[19]. Our sample includes far more Trump-supporting Republicans than typically participate in depolarization initiatives. Almost a quarter (23.4%) of respondents said they do not believe Joe Biden legitimately won the 2020 Presidential Election. Our intervention did not simply "preach to the choir" of liberals predisposed to depolarization messaging.

We also show balance between assignment groups on pre-treatment covariates in Table 9.

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

Table 7: Summary statistics (main outcome survey respondents)

Statistic	N	Mean	St. Dev.	Min	Max
Republican	3,186	0.436	0.496	0	1
Rejects 2020 election	3,185	0.234	0.423	0	1
Age	3,186	41.200	13.500	18	83
Nonwhite	3,186	0.259	0.438	0	1
College degree	3,186	0.578	0.494	0	1
Male	3,186	0.425	0.494	0	1
Pre-treatment meta-perceptions	3,180	25.600	27.300	0.000	100.000
Pre-treatment affective polarization	3,186	51.000	30.900	−100	100

Table 8: Summary statistics (followup survey respondents)

Statistic	N	Mean	St. Dev.	Min	Max
Republican	2,190	0.418	0.493	0	1
Rejects 2020 election	2,189	0.225	0.418	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

Table 9: Balance Between Treatment and Control

	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 10: Treatment and Control Groups Comparable Across Demographic and Attitudinal Covariates

7.4 Expanded summary of treatment event

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

> Prompt: This (attached) is a transcript of a public town hall event. Summarize the topics and arguments that were discussed. Do not mention the names of the speakers

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

tual 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 Regression tables for plotted models

7.5.1 Main outcomes, attitudinal

Table 11: ITT and CACE main effects coefficients, attitudinal outcomes

term	Outcome	Estimator	estimate	std.error	conf.low	conf.high	p.value
assignment	SUP	ITT	0.031	0.035	-0.037	0.099	0.373
assignment	Aff. Pol.	ITT	-0.005	0.018	-0.040	0.030	0.771
assignment	Predictions	ITT	-0.072	0.031	-0.132	-0.011	0.020
assignment	Metaperceptions	ITT	-0.125	0.030	-0.184	-0.065	0.00004
comply	SUP	CACE	0.050	0.056	-0.060	0.159	0.373
comply	Aff. Pol.	CACE	-0.008	0.028	-0.064	0.047	0.771
comply	Predictions	CACE	-0.114	0.049	-0.211	-0.018	0.020
comply	Metaperceptions	CACE	-0.199	0.048	-0.294	-0.104	0.00004

7.5.2 Main outcomes, behavioral

Table 12: ITT and CACE main effects coefficients, behavioral outcomes

term	Outcome	Estimator	estimate	std.error	conf.low	conf.high	p.value
assignment	Pledge 1	ITT	-0.017	0.017	-0.050	0.016	0.316
assignment	Pledge 2	ITT	-0.007	0.013	-0.033	0.019	0.587
assignment	Poll Worker	ITT	0.006	0.016	-0.026	0.037	0.722
assignment	Event Interest	ITT	0.085	0.017	0.050	0.119	0.00000
assignment	Behavior Index	ITT	0.016	0.011	-0.005	0.037	0.125
comply	Pledge 1	CACE	-0.027	0.027	-0.080	0.026	0.316
comply	Pledge 2	CACE	-0.012	0.021	-0.053	0.030	0.587
comply	Poll Worker	CACE	0.009	0.026	-0.041	0.059	0.722
comply	Event Interest	CACE	0.135	0.028	0.081	0.190	0.00000
comply	Behavior Index	CACE	0.026	0.017	-0.007	0.060	0.125

7.5.3 Followup Survey Experiment Models

Table 13: 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 14: 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 15: Preference for depolarization video

	(1)	(2)	(3)	(4)
Bridging video	−0.012	0.126*	0.126*	0.069
	(0.064)	(0.057)	(0.056)	(0.043)
Republican			0.082	
			(0.063)	
Bridging video*Republican			−0.138	
			(0.086)	
Pretreatment Metas				0.058+
				(0.032)
Bridging * Metas				−0.090*
				(0.043)
Num.Obs.	912	1269	2181	2181
R2	0.026	0.030	0.025	0.026
Std.Errors	IID	IID	IID	IID

+ p \num{¡ 0.1}, * p \num{¡ 0.05}, ** p \num{¡ 0.01}, *** p \num{¡ 0.001}
Models include demographic covariates

Table 16: Preference for depolarization video, binary outcome

	Republicans	Democrats	Party int.	Pretreatment int.
Bridging video	−0.106***	−0.021	−0.022	−0.057**
	(0.032)	(0.026)	(0.027)	(0.020)
Republican			0.061*	
			(0.030)	
Bridging video*Republican			−0.085*	
			(0.041)	
Pretreatment Metaperceptions				0.026+
				(0.015)
Bridging * Metaperceptions				−0.009
				(0.020)
Num.Obs.	912	1269	2181	2181
R2	0.044	0.047	0.044	0.043
Std.Errors	IID	IID	IID	IID

The outcome for these models is a binary indicator of whether a respondent stayed on the video page for at least 20 seconds. This indicates whether the respondent watched any of the video provided or whether they simply clicked through to end the survey.
+ p \num{¡ 0.1}, * p \num{¡ 0.05}, ** p \num{¡ 0.01}, *** p \num{¡ 0.001}
Models include demographic covariates

7.5.4 Full model (with covariates) for meta-perceptions

Table 17: Full ITT and CACE model outputs for standardized meta-perceptions

	ITT	CACE
(Intercept)	−0.590***	−0.601***
	(0.085)	(0.084)
Assignment	−0.125***	
	(0.030)	
Age	0.000	0.000
	(0.001)	(0.001)
raceBlack	0.077	0.095
	(0.071)	(0.071)
raceHispanic/Latino	0.099	0.114
	(0.080)	(0.080)
raceOther	−0.058	−0.045
	(0.146)	(0.148)
raceWhite	0.041	0.047
	(0.058)	(0.058)
College	−0.057+	−0.057+
	(0.032)	(0.032)
partyRepublican	0.028	0.018
	(0.032)	(0.032)
Male	−0.004	−0.005
	(0.031)	(0.031)
geoblocEastern	−0.010	−0.008
	(0.039)	(0.039)
geoblocFlorida	−0.057	−0.054
	(0.059)	(0.059)
geoblocMountain	−0.096	−0.089
	(0.066)	(0.067)
geoblocPacific/Alaska/Hawaii	0.046	0.052
	(0.049)	(0.049)
Pre-treatment meta-perceptions	0.017***	0.017***
	(0.001)	(0.001)
Pre-treatment aff. pol.	0.004***	0.004***
	(0.001)	(0.001)
Compliance		−0.199***
		(0.048)
Num.Obs.	3180	3180
R2	0.272	0.274
AIC	8047.1	8042.4
RMSE	0.85	0.85

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

7.6 Difference-in-Means Models without Covariate Adjustment

Table 18: Unadjusted difference-in-means estimates (ITT)

	(1)
Undemocratic Practices — ITT	
Assignment	0.034
	(0.035)
Affective Polarization — ITT	
Assignment	0.007
	(0.035)
Predictions — ITT	
Assignment	-0.066+
	(0.035)
Metaperceptions — ITT	
Assignment	-0.120***
	(0.035)
Pledge 1 — ITT	
Assignment	-0.018
	(0.017)
Pledge 2 — ITT	
Assignment	-0.007
	(0.013)
Poll Worker — ITT	
Assignment	0.007
	(0.016)
Event Interest — ITT	
Assignment	0.090***
	(0.018)
Behavior Index — ITT	
Assignment	0.018
	(0.011)

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001
Models include assignment block dummies

Table 19: Unadjusted difference-in-means estimates (CACE)

	(1)
Undemocratic Practices — CACE	
Comply	0.054
	(0.056)
Affective Polarization — CACE	
Comply	0.011
	(0.057)
Predictions — CACE	
Comply	-0.106+
	(0.056)
Metaperceptions — CACE	
Comply	-0.191***
	(0.056)
Pledge 1 — CACE	
Comply	-0.029
	(0.027)
Pledge 2 — CACE	
Comply	-0.012
	(0.021)
Poll Worker — CACE	
Comply	0.012
	(0.026)
Event Interest — CACE	
Comply	0.143***
	(0.028)
Behavior Index — CACE	
Comply	0.028
	(0.017)

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models include assignment block dummies

7.7 Further Compliance Analyses

7.7.1 Predictors of treatment event attendance

Table 20: 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

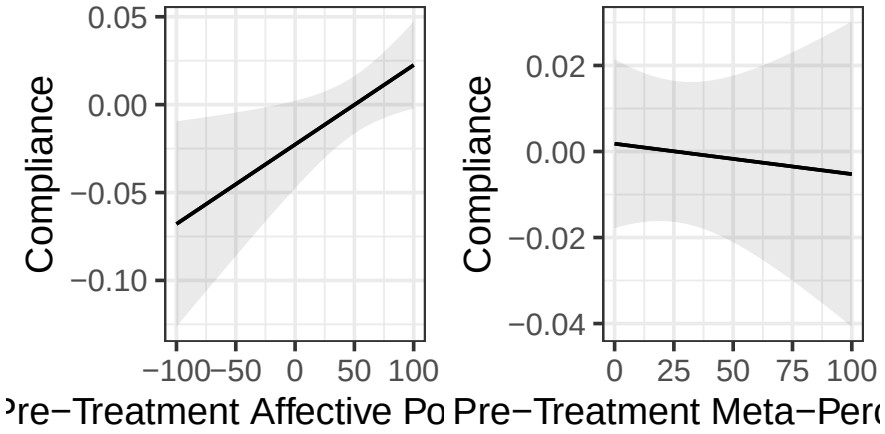


Figure 7: Continuous Pre-Treatment Affective Polarization and Meta-Perception Measures Do Not Strongly Predict Differential Compliance

7.7.2 "Super-compliance" in treatment event uptake

As shown in Figure 10, about 15% of the treatment group watched the treatment event for at least 45 minutes and 8% for at least 60 minutes. This "super-compliance" was particularly common among Black respondents — of note, the two guest speakers at the treatment event were both Black political leaders. Following the conclusion of the event, we received messages from respondents expressing their enjoyment of the event.

7.7.3 Alternative measures of compliance

Our paper has defined treatment compliance as attending the event for the 30 minutes required to receive payment. However, we could alternatively restrict our definition of compliance to event attendees who also answered both attention checks correctly or to those

Figure 8: Partisan Attendance Gap by Region

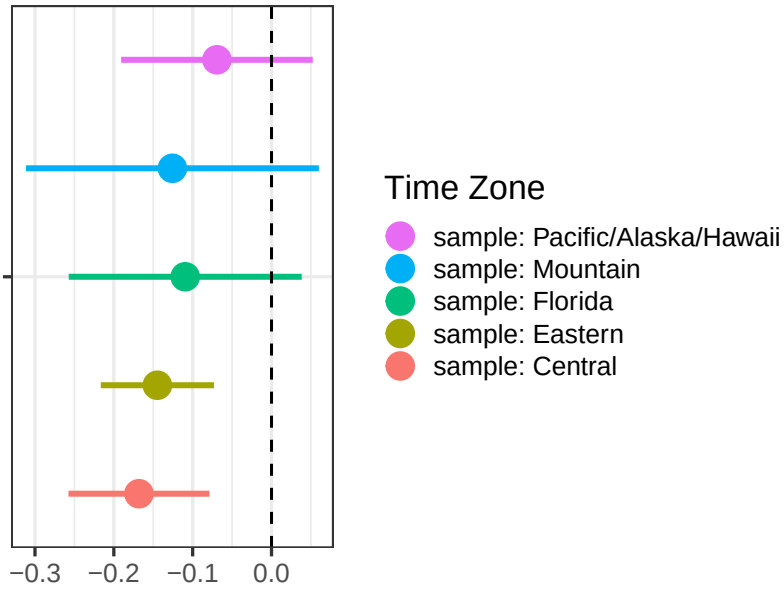
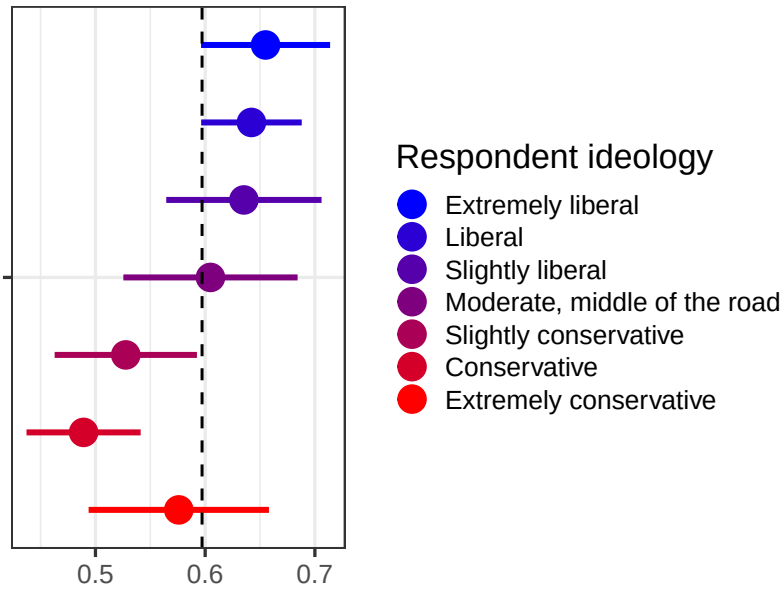


Figure 9: Attendance by ideology



who attended the live event only. Using these compliance measures to estimate the complier average causal effect (CACE) would assume that attendees who failed attention checks and recorded-event attendees, respectively, did not really receive the full treatment. Below, we show how the first stage gaps in compliance rates and the second stage CACE estimates change across more restrictive compliance measures.

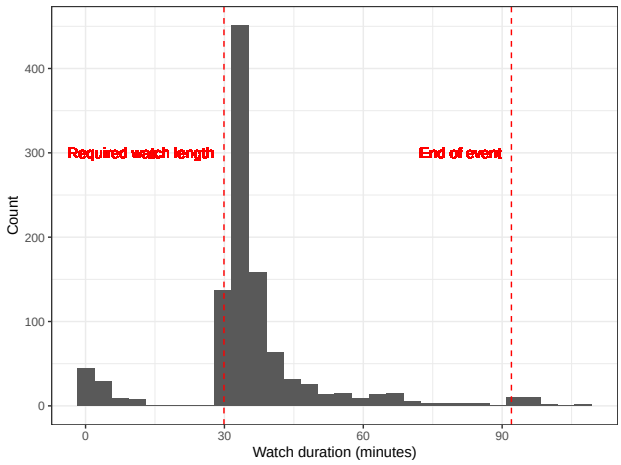


Figure 10: Histogram of event watch duration

Table 21: Party and aff. pol. compliance rate differences across different measures of compliance

	Watched	Watched and answered correctly	Attended event live only
Assignment	0.617***	0.585***	0.309***
	(0.027)	(0.027)	(0.024)
Affective polarization	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Republican	0.000	0.000	0.000
	(0.017)	(0.017)	(0.016)
Assignment:Republican	−0.132***	−0.137***	−0.119***
	(0.024)	(0.024)	(0.022)
Assignment:Aff. Pol.	0.001	0.001+	0.000
	(0.000)	(0.000)	(0.000)
Num.Obs.	3461	3461	3461
R2	0.430	0.400	0.179

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

Table 22: 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

7.8 Additional Heterogeneous Effect Analyses

7.8.1 Heterogeneous effects, main experiment

Table 23 tests for treatment effect heterogeneity by the timing of participants viewing the event. Both models use standardized meta-perceptions as the outcome of interest. In Model 1, we compare participants who watched the treatment event live to those who watched the recording of the event and find no difference in treatment effects. In Model 2, we test for an interaction between treatment assignment and the timing of respondents' completing the outcome survey, which would indicate a rapid decay of the treatment effect. We find no significant interaction.

Table 23: Heterogeneous effects by live vs recorded watch and by survey delay

	(1)	(2)
Watched event live	−0.019	
	(0.053)	
Assignment		−0.130**
		(0.050)
Survey completion time		0.001*
		(0.001)
Assignment*Survey completion time		0.000
		(0.001)
Num.Obs.	998	3180
R2	0.242	0.274
Std.Errors	IID	IID

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

Model 1 limited to treatment group compliers. Model 2 uses full sample. Both models include covariates and block dummies

7.8.2 Heterogeneous effects, followup survey experiment

Figure 11 plots the coefficient of *the interaction term Treatment*Prior Meta perceptions* in the followup survey experiment. In this figure, a positive coefficient means that more negative meta-perceptions of the opposite party are associated with a more positive treatment effect estimate. We use respondents' answers to the meta-perceptions questions at the beginning of the followup survey to estimate this interaction. As in Figure 6 in the main text, we compare the elite and voter corrections to the pure control condition. We see no significant heterogeneity by respondent meta-perceptions among Republicans for any of the four outcomes. Among Democrats, we do find that higher meta-perceptions (of Republicans holding antidemocratic attitudes) interact with both correction treatments *negatively* for predictions of opponent norms violations and *positively* for support for own-party norms violations. These results are difficult to interpret, as the main non-interacted effect estimates were mostly non-significant. We can tentatively say, however, that at least the elite correction was particularly effective among Democrats with more negative prior expectations about Republicans.

Table 24 looks for differences in preference for the bridging video not by partisanship, but instead by respondents' pre-treatment affective polarization and meta-perceptions. We find that participants with higher pre-treatment meta-perceptions (that is, more negative estimates of the opposing party's beliefs) watched the bridging video for a shorter duration, relative to the co-partisan messaging placebo.

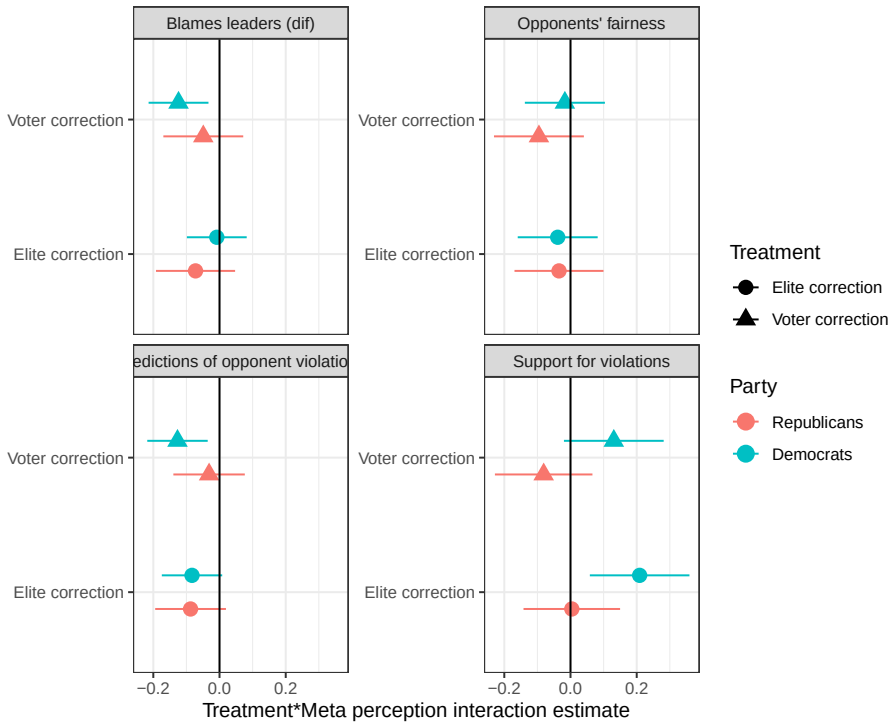


Figure 11: Heterogeneous effects by meta-perceptions as measured at the beginning of the followup survey

Table 24: 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

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 25 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 25: 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 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.

7.11 Follow-up attrition

As noted in the paper, the survey experiments embedded in the follow-up study featured new randomizations. Attrition is thus not problematic for these new experiments, only for analyzing the persistence of the original treatment event's meta-perceptions correction. Below, we test for potentially problematic attrition by the time of the follow-up.

Table 26: Followup attrition is Not Predicted by Treatment Assignment

	Followup attrition	Followup attrition covariates
Assignment	0.007	0.007
	(0.016)	(0.016)
Num.Obs.	3461	3453
R2	0.000	0.058
Std.Errors	IID	IID

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

Second model includes demographic covariates

We find no effect of assignment to the treatment event on likelihood of completing the followup survey. As in the main outcome survey, non-response is independent of treatment assignment.

7.12 Follow-Up Experimental Materials

Table 27: 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?"	