Rethinking 'Them': Challenging Out-Group Stereotypes in Backsliding Democracies

Preprint

Gencer, Alper Sukru

2026-05-02

Word count: 5558

Abstract:

Affective polarization carries particular force in backsliding democracies, where citizens often exaggerate the extremity of their political opponents and may distrust corrective information. This study asks whether targeted, issue-specific stereotype correction can moderate those exaggerated beliefs in Türkiye, and whether its downstream effects vary by intervention format and audience. I first develop a data-light targeting procedure that uses existing opinion surveys to identify policy domains on which out-group extremity is especially likely to be overstated. I then field a preregistered online survey experiment comparing two interventions: accuracy feedback about the true distribution of out-group views, and exposure to simulated out-group conversations expressing moderate positions. Accuracy feedback leads respondents in the full sample to see the political out-group as less extreme, with larger corrective point estimates among opposition supporters, whereas conversation exposure yields weaker average belief updating. Those shifts in belief do not translate uniformly into warmer intergroup attitudes. The affective responses are more heterogeneous and often less precisely estimated, with subgroup patterns suggestive of more favorable relational responses under conversation exposure among opposition supporters, who also begin with the largest baseline misperceptions, and less favorable affective responses to accuracy feedback among government supporters. Taken together, the findings suggest that belief correction remains feasible under democratic backsliding, but that its affective consequences depend on delivery format, audience characteristics, and the credibility of the environment in which the message is received.

Keywords: affective polarization; democratic backsliding; out-group stereotypes; survey experiment; belief updating; Türkiye

1 Introduction

Affective polarization carries particular weight in backsliding democracies, where incumbents routinely frame politics as a form of existential struggle and institutional constraints on intergroup hostility have weakened (Iyengar, Sood, and Lelkes 2012; Svolik 2019; Druckman et al. 2022). Citizens often overestimate how extreme political out-groups are, and such misperceptions predict social distance as well as support for undemocratic behavior (Ahler and Sood 2018; Voelkel et al. 2023). In consolidated democracies, efforts to correct these misperceptions can sometimes reduce hostility, but their effects vary across formats and outcomes (Druckman et al. 2023; Hartman et al. 2022; Voelkel et al. 2023).

This paper argues that, in backsliding settings, belief correction and affective depolarization can come apart because credibility and identity threat shape how corrective messages are received. In captured media environments, source legitimacy is itself politicized, so recipients may update on facts while also revising beliefs about the communicator's intent (Rozenas and Stukal 2019; Barberá 2020; Shirikov 2024). Direct corrective prompts may therefore be filtered through motivated reasoning and source distrust, making rejection or affective backlash plausible even when factual updating occurs (Taber and Lodge 2006; Kahan 2013; Vegetti and Mancosu 2020; Nyhan 2021). Rather than attempting to revise views of the out-group in the abstract, I focus on stereotype-prone issues and ask whether issue-specific evidence of out-group moderation yields more favorable results by reducing identity threat. To do so, I first use a representativeness-discounted population density (RDPD) procedure to identify stereotype-prone policy issues, and then test two formats that convey the same substantive signal of out-group moderation: an accuracy-feedback correction that reports survey distributions and a conversation-exposure intervention that presents simulated out-group conversations.

To evaluate these questions, I field a preregistered online survey experiment in Türkiye. Accuracy feedback shifts perceived out-group extremity toward moderation in the full sample, while conversation exposure produces weaker and more heterogeneous belief updating. These belief changes do not translate uniformly into warmer intergroup attitudes. The affective effects are smaller, more heterogeneous, and often imprecisely estimated, with directional subgroup patterns that differ across the two interventions. That divergence matters because it suggests that, in backsliding democracies, the central question is not only whether corrective information changes beliefs, but whether a given format can do so without activating identity threat or source distrust.

The paper contributes both a targeting tool and new evidence on the portability of stereotype-correction interventions to backsliding democracies, which remain comparatively understudied (Iyengar, Sood, and Lelkes 2012; Iyengar and Westwood 2015; Druckman et al. 2022; Voelkel et al. 2023). The RDPD procedure offers a data-light way to identify stereotype-prone issues from standard opinion-survey marginals, extending representativeness-based accounts of stereotyping into the domain of intervention targeting (Bordalo et al. 2016). More broadly, the findings show that belief correction and affective depolarization can come apart in polarized information environments. Accuracy feedback is the stronger tool for correcting exaggerated beliefs, but its affective consequences are audience-dependent. Conversation exposure yields weaker average belief correction, yet its most favorable relational estimates appear among opposition supporters. Taken together, these patterns suggest that, in backsliding settings, whether correction reduces hostility depends not only on informational content, but also on credibility, delivery format, and audience.

2 Theory: Stereotype Correction Under Democratic Backsliding

Affective polarization reflects how citizens feel about political groups, not only what they believe about policies. It often appears as social distance, dislike, and unwillingness to engage with out-group members in everyday life (Tajfel et al. 1979; Iyengar, Sood, and Lelkes 2012; Iyengar and Westwood 2015). In backsliding democracies, leaders and aligned media regularly frame politics as a struggle for survival. That framing can raise the stakes of intergroup conflict and weaken institutional constraints on hostile behavior (McCoy, Rahman, and Somer 2018; Svolik 2019; Gessler and Wunsch 2023). These conditions make stereotype correction harder, because the same message can be processed as information by some audiences and as threat by others.

2.1 Why captured media environments amplify extremity stereotypes

Citizens often overestimate how many out-group members hold extreme positions. Two forces make this pattern especially likely under backsliding. First, captured and polarized media systems skew what people see. Regime-aligned outlets amplify conflictual and extreme content about opponents, while opposition voices face constraints, marginalization, or selective exposure channels (Rozenas and Stukal 2019; Barberá 2020). This shifts the sample of out-group signals that enters citizens' attention.

Second, citizens generalize from salient cases. When extreme out-group positions are vivid, people infer that those positions are common. This inference is consistent with representativeness-based stereotyping. Citizens overweight types that fit a salient prototype and underweight base rates (Bordalo et al. 2016; Lees and Cikara 2020). As a result, misperceptions should be largest on issues where (i) extreme positions are easy to recall or widely circulated, but (ii) the true out-group distribution places relatively little mass at the extreme. This logic motivates a targeted approach. Instead of correcting beliefs on any random issue, I focus on issues where stereotyping is most likely. The RDPD procedure operationalizes this targeting strategy using existing opinion surveys. This targeting also serves a design purpose. By correcting beliefs through concrete, domain-specific policy distributions rather than asking respondents to evaluate out-group extremity in general, the intervention aims to lower identity salience and reduce opportunities for motivated reasoning, even though it may not eliminate credibility-based defensiveness in captured media environments.

2.2 Why belief updating may not translate into warmer attitudes under backsliding

Belief correction need not reduce affective polarization because factual beliefs about out-group extremity and relational attitudes toward the out-group are distinct objects. Corrections can shift perceived distributions while leaving threat perceptions, identity-based dislike, or moralized judgments unchanged.

One reason is motivated reasoning. Individuals can accept some factual content while defending prior affective orientations, especially when the information carries implicatures about blame, legitimacy, or status (Taber and Lodge 2006; Kahan 2013; Jenke 2023). In polarized and partially captured information environments, this separation can be stronger because credibility is itself politicized. The same corrective message may be processed as neutral evidence by some audiences and as hostile signaling by others, even when the factual content is accepted (Nyhan 2021).

In captured and polarized information environments, credibility is politicized. The same correction can function as information for some audiences and as identity threat for others (Shirikov 2024). This creates a conditional belief-to-affect link. When recipients view the message as credible and non-threatening, belief updating is more likely to translate into lower social distance. When recipients experience the message as hostile auditing or an attack on group standing, belief updating may not reduce social distance and can coincide with greater distance. This logic motivates treating perceived legitimacy and identity threat as moderators of downstream affect.

2.3 Intervention format and audience psychology

The two interventions tested here differ in how they present out-group moderation. The accuracy-feedback correction discloses true out-group survey distributions on the targeted issues. This format directly addresses base-rate neglect. It should move posterior beliefs when citizens attend to the information and treat the source as minimally credible (Druckman et al. 2023; Voelkel et al. 2023). At the same time, in a captured media environment it can resemble fact-checking by hostile elites. That interpretation can trigger identity defense among regime-aligned citizens (Shirikov 2024).

The conversation exposure presents simulated out-group conversations that express moderate positions using familiar discourse. Narratives can make moderation vivid and socially legible without explicitly confronting citizens with a corrective verdict (Shiller 2017; Bursztyn, González, and Yanagizawa-Drott 2020). This may reduce defensiveness and facilitate affective change among audiences that are open to interpersonal imagination. However, conversation exposure may have weaker effects on factual beliefs if respondents treat it as anecdotal or unrepresentative.

Note that the two treatments are matched on topic and direction of the signal (out-group moderation), but they are not informationally equivalent: accuracy feedback provides an explicit base-rate statistic, while

conversation exposure provides anecdotal social evidence that may be processed as less representative. In short, accuracy feedback should yield stronger belief correction but carries higher affective backlash risk, whereas conversation exposure should yield weaker belief correction but is more affect-safe for relational outcomes.

Figure 1 summarizes the core causal logic. The treatments can shift posterior beliefs about out-group extremity. Those belief changes can reduce social distance, but identity threat and source distrust can weaken or reverse the downstream affective response.

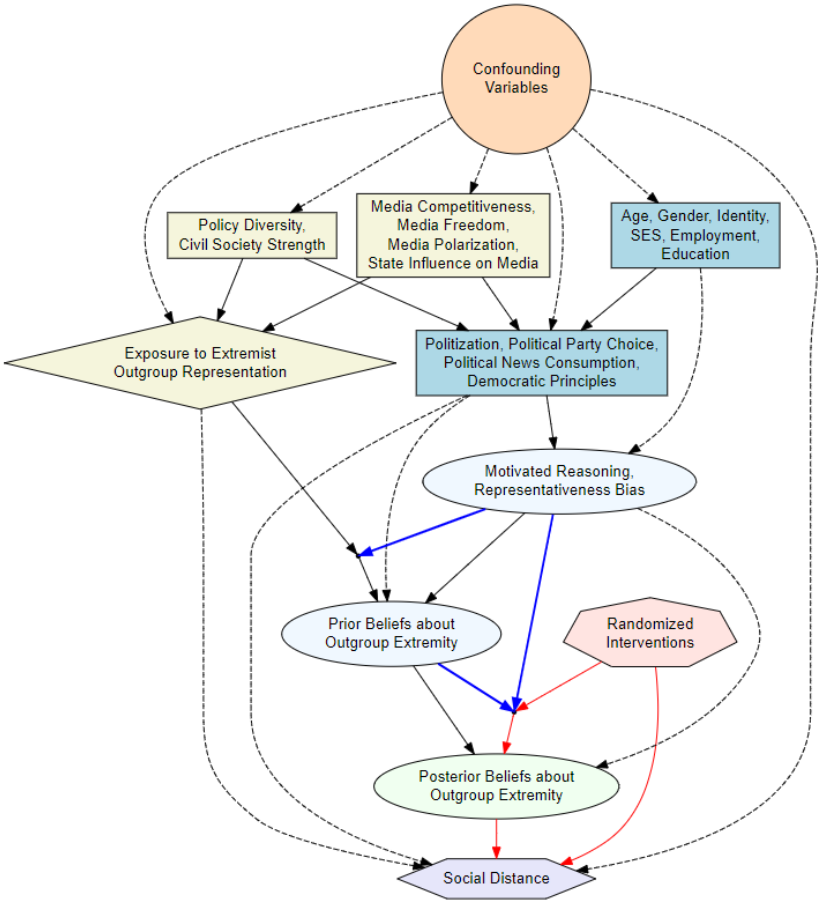

Figure 1: Directed acyclic graph of social distance in backsliding democracies. Blue arrows indicate moderating effects.

2.4 Hypotheses

My preregistration specifies two confirmatory outcome families: (i) perceived overall out-group extremity and (ii) affective polarization toward the out-group (social distance and negative trait associations). The hypotheses test whether the interventions reduce perceived out-group extremity and whether any belief change translates into lower affective polarization.

Backsliding settings with captured media politicize credibility, creating a format trade-off. Accuracy feedback should produce larger belief updates because it provides explicit base-rate information, but it may be affect-risky if recipients experience numerical correction as hostile auditing or identity threat. Conversation exposure should be more affect-safe for relational outcomes because it presents out-group moderation in a socially legible, less confrontational form, even if it yields weaker average belief updating.

H1 (Affective polarization). Relative to placebo, both interventions should reduce affective polarization toward the out-group, measured by social distance and negative trait associations. However, because credibility is politicized under backsliding, I expect affective responses to be more heterogeneous than in consolidated democracies. In particular, accuracy feedback is more likely to generate affective backlash in some audiences, while conversation exposure should be more affect-safe on relational outcomes.

H2 (Belief correction). Relative to placebo, both accuracy feedback and conversation exposure should reduce perceived out-group extremity on the targeted issues and shift posterior beliefs toward the benchmark distributions implied by the Konda surveys. Because accuracy feedback provides explicit distributional statistics, it should produce larger belief updating than conversation exposure on average.

H3 (Cognitive effort/Duration). The preregistration also specifies response duration (log time-on-page) for two post-treatment pages as auxiliary outcomes capturing engagement with the material. I do not treat the substantive responses on those pages as confirmatory outcomes.

Scope conditions and preregistered heterogeneity. The preregistration anticipates heterogeneous effects by baseline priors, partisan attachment, and education. Substantively, the clearest implication concerns baseline priors about out-group extremity. I preregistered a non-monotonic responsiveness expectation. Interventions should be most productive among respondents with moderate misperceptions, where there is meaningful scope to update and where corrections are less likely to be filtered through identity defense. By contrast, respondents with very high prior extremity beliefs may be less responsive despite greater nominal scope for updating, because motivated reasoning and selective exposure can make counter-stereotypical information easier to dismiss. At the other end, when priors are already relatively optimistic or close to benchmark values, additional correction has limited room to move beliefs and may even provoke reactance if it is experienced as an evaluative audit rather than neutral information. I also preregister education and partisan affiliation strength as moderators, expecting larger responses among more educated respondents and among those with weaker partisan attachment.

Exploratory implication (information environment). Motivated by the backsliding framework, I also explore whether treatment responses vary by respondents' self-reported primary news source, with particular attention to mainstream channels. If mainstream exposure proxies regime-aligned credibility cues, backlash-type affective responses should be more likely among regime-aligned respondents embedded in mainstream media, even when belief updating occurs. This analysis was not preregistered, so I present it as descriptive and hypothesis-generating.

3 Research design

3.1 Türkiye as a case for stereotype correction under democratic backsliding

To test these hypotheses, I conducted an online survey experiment in Türkiye, a hard test for stereotype correction under democratic backsliding. Contemporary Turkish politics is organized around a sharp cleavage between supporters of President Recep Tayyip Erdoğan's governing coalition and its opponents, a divide that intersects with salient ethnic, religious, and cultural identities (including Kurds, Alevis, Armenians, and secularists) and sustains durable identity-based conflict (Somer 2019; McCoy, Rahman, and Somer 2018; Erdoğan and Onar 2020; Orhan 2022; İmamoğlu 2022). Over the past fifteen years, democratic institutions have eroded through declining judicial independence, constraints on press freedom, and weakened checks and balances, conditions that intensify polarization and raise the perceived stakes of intergroup conflict (Esen 2021; Moral 2017; McCoy, Rahman, and Somer 2018).

These political changes also shape the information environment in ways that are directly relevant for intervention design. State-aligned mainstream media and polarized outlets amplify conflictual narratives and selectively portray political opponents, while opposition perspectives face marginalization (Erdoğan and Onar 2020). Economic stressors, including high inflation, unemployment, and broader dissatisfaction, further heighten grievance and negative affect (Kutlay and Öniş 2024; Erdoğan and Onar 2020). At the same time,

social media has become a central venue for political discourse, facilitating selective exposure and reinforcing partisan hostility (Erdoğan and Onar 2020). Together, these features make Türkiye an informative setting to evaluate whether stereotype-challenging exposure can correct misperceptions about political out-groups and whether any belief change translates into reduced social distance when source trust is politicized and identity threat is salient.

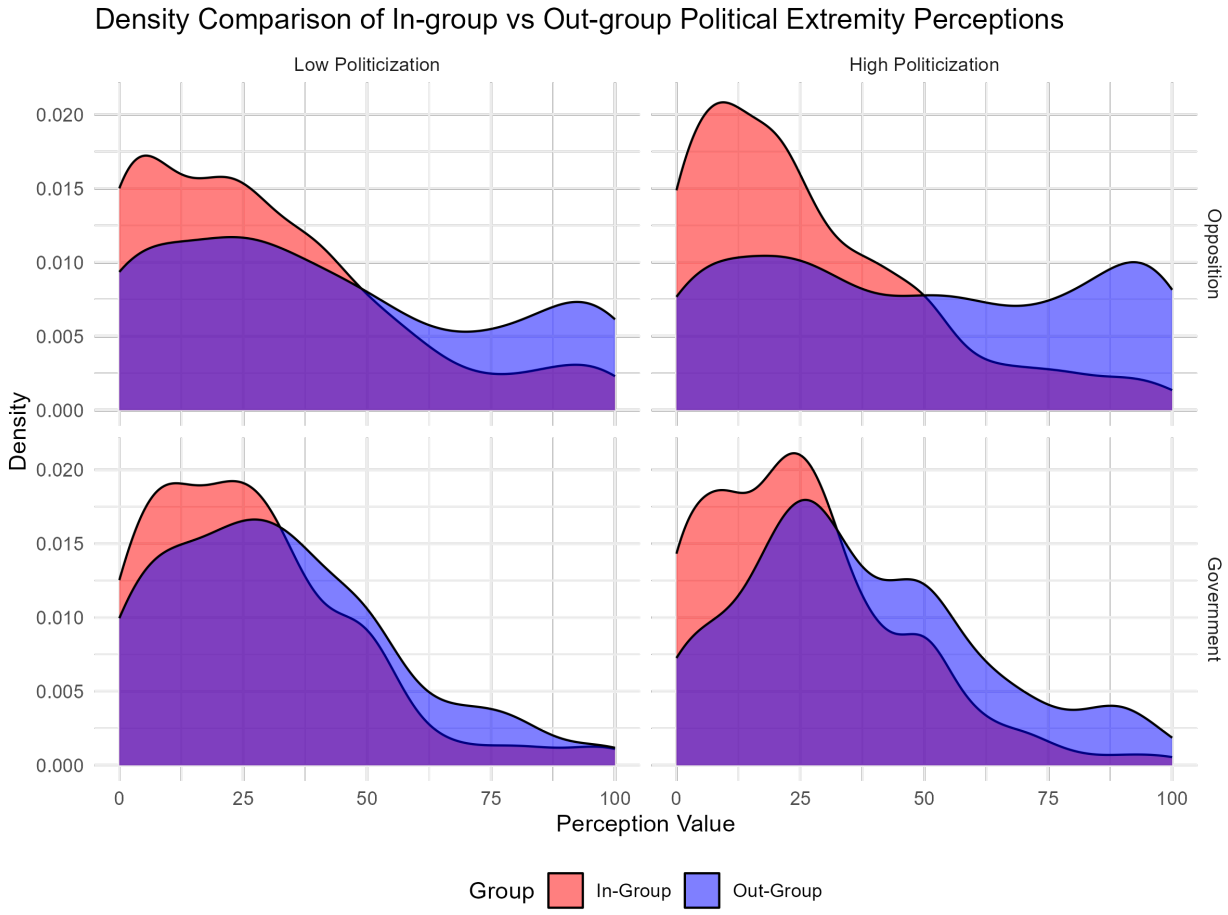

Figure 2: Comparison of In-group and Out-group Political Extremity Perceptions. (N=1263).

Before turning to treatment effects, I document baseline misperceptions and affective polarization in the study population. I focus on two descriptive patterns that motivate the main outcome families: (i) distribution of perceived out-group extremity relative to the in-group, (ii) social distance in everyday interactions, and (iii) negative trait attributions toward the out-group. (For additional descriptive evidence on perceived willingness of the out-group to use undemocratic or hostile tactics, see Appendix 1.)

Figure 2 compares perceived extremity distributions for in-groups and out-groups. Respondents consistently perceive the out-group as more extreme than the in-group, and the perceived gap is especially pronounced among opposition supporters and among more politicized respondents.

Table 1: Social Distance Questions by Partisanship

Statements: "I would feel uncomfortable if ..."	Government Supporters (%)	Opposition Supporters (%)	P-value (Chi-square)
... if my children becomes friends with the children of out-group members.	30.188	52.856	<0.001
... doing business with supporters of other party.	23.830	39.058	0.002
... if my neighbors are out-group members.	36.558	50.056	<0.001
... if my child married someone who is a supporter of the government.	33.194	43.350	<0.001

Source: Author's original survey experiment (population-weighted). Proportions are the percentages of respondents who selected "somewhat agree" or "strongly agree." Estimates use the placebo group only to describe baseline polarization (N = 434).

Table 1 reports social distance. Respondents report substantial discomfort with cross-partisan relationships in family, neighborhood, and economic life, with opposition supporters typically expressing higher discomfort, consistent with severe affective polarization.

Finally, Table 2 summarizes negative trait attributions toward the out-group. While government supporters frequently attribute negative traits to opposition supporters, opposition supporters attribute them at even higher rates to government supporters. Together, these patterns show that both belief-based and affective divisions are salient in this setting, making Türkiye an informative context for evaluating stereotype-challenging interventions under democratic backsliding.

Table 2: Negative Trait Association Questions by Partisanship

Statements: "Compared to in-group supporters, out-group supporters are more..."	**Government Supporters (%)**	**Opposition Supporters (%)**	**P-value (Chi-square)**
... close-minded.	38.947	74.601	<0.001
... immoral.	34.885	58.712	<0.001
... lazy.	49.181	63.275	<0.001
... dishonest.	45.691	64.463	<0.001
... unintelligent	47.787	71.114	<0.001
... traitorous.	25.824	42.717	0.013

Source: Author's original survey experiment (population-weighted). Proportions are the percentages of respondents who selected "somewhat agree" or "strongly agree." (N = 434).

3.2 Recruitment Procedure and Experimental Design

I recruited participants through Meta advertisements on Facebook and Instagram. The advertisement campaign and data collection took place between January 24 and January 31, targeting Turkish citizens aged 18 and older. In line with Meta's transparency procedures, I ran "social issues, elections, or politics" ad campaigns using a research organization account, registered with a Turkish phone number and bank account. To improve recruitment efficiency, I used a Meta "conversion" campaign that optimizes ad delivery using a survey-completion event. Neundorf and Öztürk (2021) shows that conversion campaigns can improve recruitment efficiency without producing observable imbalances in sample demographics. Although running multiple advertisement sets with different targeting strategies can improve balance in Meta convenience samples (Neundorf and Öztürk 2021), Meta limits demographic targeting for political ads. Consequently, I fielded the study using a single non-targeted ad set (for Instagram and Facebook advertisement images, refer to Figure 3). Participants were incentivized through automatic enrollment in a raffle with the chance to win coupons totaling 200 Turkish Lira, redeemable at two major market chains.[1]

Upon clicking the ads, participants were redirected to Qualtrics, where the survey was administered. To reduce demand effects and motivated responding, the initial consent script described the study in general terms, and participants received a full debriefing at the end of the survey. The incentive scheme was implemented through raffle entries: all participants who completed the survey were enrolled in the raffle, and correct answers to attention checks increased participants' chances of winning.[2]

After informed consent and eligibility screening (age and citizenship), participants completed a pre-treatment survey measuring demographics, news consumption, political interest, political affiliation, past voting behavior, democratic support, political knowledge, economic satisfaction, and prior beliefs regarding in-group and out-group extremity and willingness to compromise. Subsequently, Qualtrics randomly assigned participants to one of three experimental conditions: (1) Accuracy Feedback Treatment, (2) Conversation Exposure Treatment, and (3) Placebo Group. Following exposure to their respective treatments, participants answered post-treatment attention questions, reported posterior beliefs, and completed items measuring out-group perceptions and affective polarization.[3]

Ethical approval for this experiment was granted New York University's Institutional Review Board (IRB-FY2024-8112) on January 3, 2024. Additionally, a detailed pre-analysis plan was preregistered prior

[1]The ads included the following text: "Would you like to participate in this incentivized scientific research organized by New York University? Everyone who completes the 15-minute survey has a chance to win a discount voucher worth 200 TL!"

[2]The study also included a belief elicitation task. The main incentive instrument in the fielded design was raffle-entry based rather than direct performance pay. For the results of the elicited beliefs on the targeted issues, refer to the figure in Appendix 2.

[3]All questions were preregistered before the research. For a comprehensive list of all pre- and post-treatment questions, refer to the Questionnaire in Appendix 11.

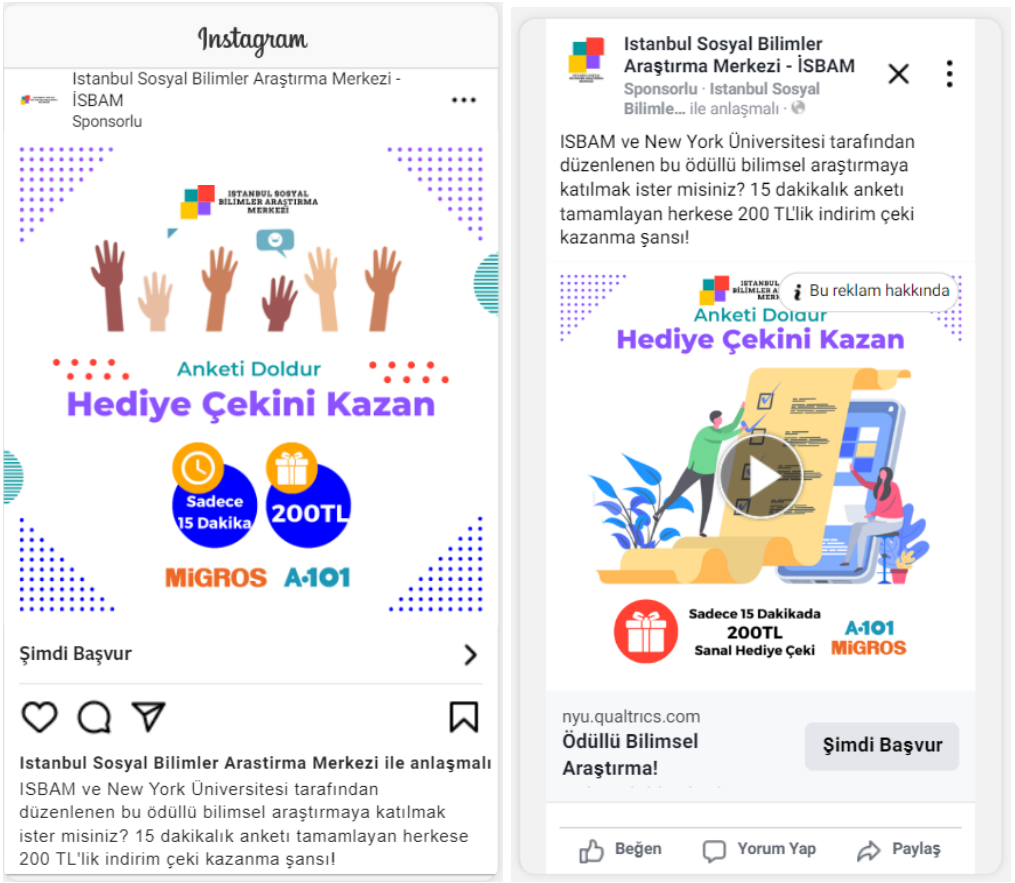

Figure 3: Meta Recruitment Ads.

to fielding the study on OSF Registries on January 22, 2024, specifying the hypotheses, research design, estimation methods, variables, and questionnaire.[4][5]

3.3 Determining Stereotype-Prone Policy Issues (RDPD targeting)

The interventions target policy issues where respondents are most likely to exaggerate out-group extremity. In backsliding settings, researchers rarely have direct measures of out-group perception errors across a wide issue set. I therefore develop a targeting procedure that relies only on standard opinion-survey marginals.

The procedure builds on representativeness-based accounts of stereotyping (Bordalo et al. 2016). The key intuition is that stereotypes are most distortionary when an extreme position is (i) diagnostic of group difference and therefore salient, but (ii) not actually common in the out-group. In that case, citizens can overweight a vivid extreme relative to the true population frequency.

Operationally, I use Konda surveys to identify divisive issues and then compute, for each issue and each extreme response category, a diagnosticity statistic. For issue k and response category $x^{extreme}$, I define a likelihood ratio that captures how diagnostic that extreme response is of out-group membership:

$$LR_k(x^{extreme}) = \frac{P(X_k = x^{extreme} \mid Out - group)}{P(X_k = x^{extreme} \mid In - group)}.$$

I then discount this diagnosticity by the out-group's true population density at that extreme response, so that the targeting score is highest when an extreme response is strongly diagnostic but rare in the out-group. I rank issues by this representativeness-discounted score and select a small set for treatment content (full details and equations are reported in Appendix 2.).

I preregistered the issue-selection rule and, from the highest-ranked issues, selected low-threat domains that are divisive yet unlikely to trigger direct policy conflict in the survey setting. For opposition respondents, the intervention content targets misperceptions about government supporters' extremity on housing opportunities and income equality. For government respondents, it targets misperceptions about opposition supporters' extremity on infrastructure and road construction.

Figure 4 shows that, before treatment, both groups substantially overestimate the share of the out-group endorsing the extreme position on these domains in an incentivized question. These selected issues therefore provide a focused test of whether targeted evidence about out-group moderation can shift perceived out-group extremity and, in turn, affective polarization. The design evaluates whether this targeting strategy yields substantively useful intervention domains, not whether RDPD is uniquely optimal relative to alternative issue-selection rules.

3.4 Interventions Operationalization

The interventions are designed to correct exaggerated beliefs about out-group extremity on the targeted issues. Because informational treatments can work differently depending on respondents' prior beliefs and willingness to engage with evidence (Haaland, Roth, and Wohlfart 2023), I implement two formats that convey the same substantive message, out-group moderation, through different psychological channels.

First, the *accuracy-feedback* treatment provides a short, factual correction that reports true out-group response shares from Konda surveys on the selected issues (see Appendix 4). The goal is to address base-rate neglect directly by presenting clear population frequencies and thereby inducing belief updating in the direction of the true distribution (Bordalo et al. 2019; Bursztyn, González, and Yanagizawa-Drott 2020). In the survey, the text is paired with simple comprehension prompts and attention checks to ensure respondents attend to the numerical content.

Second, the *conversation exposure* treatment conveys out-group moderation through a simulated social-media exchange that uses familiar language and recognizable conversational cues (see Appendix 4). The content was constructed from a random sample of 50,000 classified Turkish political tweets, anonymized to protect privacy and reconstructed for Facebook.[6] This format is motivated by social identity and narrative

[4]The preregistered analysis plan is accessible via <https://osf.io/gdh4w>.

[5]Deviations from the preregistration are discussed in Appendix 3.

[6]The tweets are randomly sampled from political tweets on Turkish Twitter between January 2023 and June 2023. To safeguard the privacy of tweet owners, I anonymized the tweets by altering the content and introducing common internet typos.

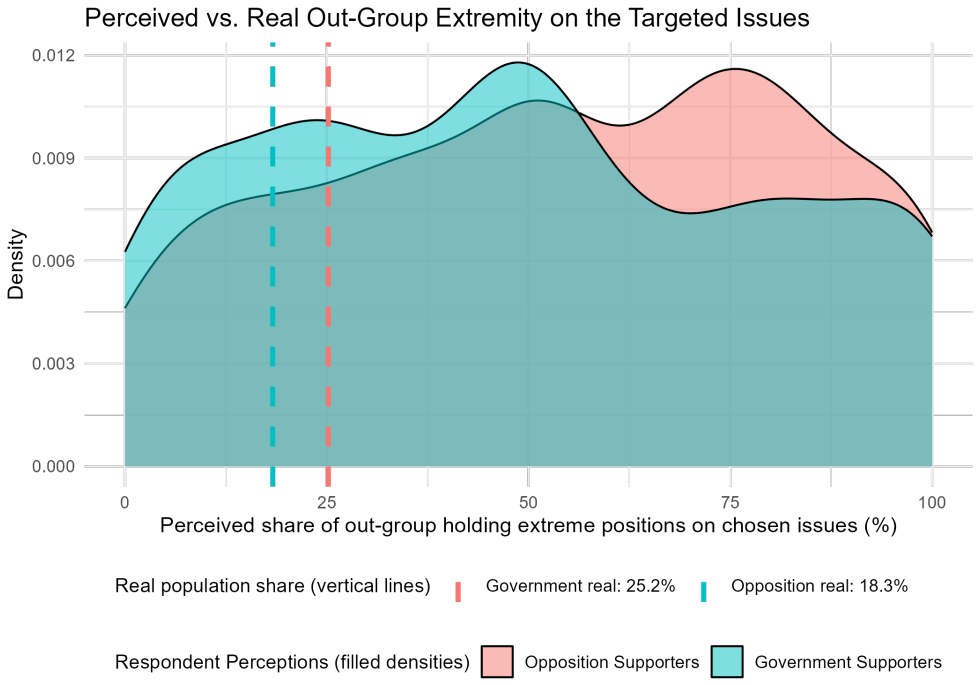

Figure 4: Elicited Beliefs (Percentage of Out-group Members Holding Extreme Positions). Dashed lines indicate true shares of extreme positions (N=1263).

mechanisms: when a correction feels like an external verdict, respondents may resist it; when moderation is presented as socially legible and relatable, it may reduce defensiveness and facilitate affective change (Turner 1975; Tajfel et al. 1979; Berger and Quinney 2004; Shiller 2017; Vives, Cikara, and FeldmanHall 2022).[7]

Both treatments are tailored to respondents' political camp. Opposition respondents receive content about government supporters' positions on housing opportunities and income equality, while government respondents receive content about opposition supporters' positions on infrastructure and road construction. The placebo group reads a neutral, non-political text of comparable length to absorb attention and provide a baseline for estimation (see Appendix 4).

3.5 Pre-Treatment and Post-Treatment Operationalization

The Qualtrics survey comprised three sections.[8] In the first section, participants completed a pre-treatment survey capturing demographic information (age, gender, income, education, ethnicity, religion), political interest, news consumption, past party choice, partisan identification strength, democratic support, and economic satisfaction, and they completed an attention check.

A central pre-treatment measure captures beliefs about the distribution of out-group extremity. Participants allocated 100 hypothetical individuals from each relevant group into three categories: *politically compromising*, *compromising in some matters*, and *politically uncompromising and extreme*. This allocation task is designed to elicit perceived belief distributions rather than a single-point stereotype and follows established approaches for measuring distributional beliefs in surveys (Leemann, Stoetzer, and Traunmüller 2021).

I incorporated design features to mitigate choice set effects and order effects that can inflate perceived out-group extremity. First, to reduce exaggeration when the political out-group is evaluated in isolation, respon-

[7] Related discussions of identity, persuasion, and dynamics of belief change are summarized in @thaler2020good and @zimmermann2020dynamics.

[8] All questions were preregistered before the experiment. For a detailed list of all pre- and post-treatment questions, see the Questionnaire in Appendix 11.

dents also completed comparative allocation questions about their in-group and about immigrants/refugees. Because immigrants/refugees are often a more stigmatized group, this comparative set reduces the tendency to default to high extremity ratings for the political out-group. Second, to limit anchoring and sequencing artifacts, I used immigrant perceptions as an initial anchor and then randomized the order of out-group and in-group questions (Maniadis, Tufano, and List 2014; Chaudoin, Gaines, and Livny 2021).[9] In addition, participants completed an incentivized belief-elicitation item about specific out-group policy positions, and correct estimates increased raffle odds, encouraging accuracy (Schotter and Trevino 2014).

In the second phase, Qualtrics randomly assigned participants to one of three conditions: the *accuracy-feedback* treatment, the *conversation exposure* treatment, or a *placebo* text.[10][11] Post-treatment, participants answered attention-check questions and briefly described their exposure.

In the final section, participants completed post-treatment measures of outcomes and candidate mediators. These included posterior beliefs about out-groups and in-groups, negative trait associations (Voelkel et al. 2023), social distance (Iyengar, Sood, and Lelkes 2012; Klar, Krupnikov, and Ryan 2018), and evaluation tasks measuring perceived out-group extremism. Outcome blocks were randomized to minimize order effects and cross-measure contamination (Chaudoin, Gaines, and Livny 2021).

3.6 Sample Description

Meta advertisements reached 127,000 individuals, yielding approximately 2125 completed surveys. After removing duplicate entries, incomplete submissions, and respondents who later decided not to share their responses, the final Meta convenience sample comprised 1263 participants.[12] Table 3 reports descriptive statistics for the analytic sample.[13]

The sample is broadly balanced on key demographics, including gender and education. It is predominantly ethnic Turkish (86), and 35 identify as either a non-Turkish ethnicity or a non-Sunni Muslim. Political participation is high: 92 report voting in the last general election, and 41 report supporting the ruling coalition. Random assignment yielded treatment shares of 32 in the *accuracy-feedback* condition, 34 in the *conversation exposure* condition, and 34 in the *placebo* condition.

Balance checks across treatment groups show no statistically meaningful differences for almost all covariates. The only exception is full democratic support, which is lower in the placebo group than in the two intervention groups, while partial democratic support does not vary across conditions.[14] While the sample is a Meta convenience sample and therefore not nationally representative, its composition is consistent with other Meta-recruited samples in Türkiye.[15] All descriptive statistics use post-stratification population weights to align the sample with Turkish population benchmarks. All treatment effect estimates reported in the main text are unweighted intent-to-treat estimates; weighted estimates are reported as robustness checks in Appendix 8.

[9] An anchor serves as a reference category and can increase variance and nuance in subsequent responses by discouraging mechanical use of extreme response categories [@maniadis2014one; @chaudoin2021survey].

[10] Participants could exit the survey before treatment exposure, allowing uninterested respondents to discontinue participation.

[11] Prior to the main study, I conducted an in-person focus group and a pilot survey via Meta recruitment to evaluate treatment materials and assess response and completion rates.

[12] The experiment is powered to detect an effect size of 0.5 standard deviations with a targeted sample of 700 observations. These calculations do not imply comparable power for differential effects across partisan subgroups, because heterogeneity tests rely on much smaller within-partisan treatment cells. As a rule of thumb, splitting the sample into government and opposition supporters roughly doubles the minimum detectable difference in treatment effects across groups, holding the total sample size fixed. Accordingly, null findings on partisan heterogeneity should be interpreted as imprecise rather than as evidence of no partisan differences, and I therefore emphasize effect sizes and confidence intervals when discussing subgroup patterns. See Appendix 5 for the power analysis.

[13] All variables in the summary table are binary.

[14] See Appendix 6 for the balance table with covariates and p-values from F-tests across groups.

[15] For discussion of Turkish Meta convenience samples, see @neundorf2021recruiting. Appendix 7 compares the sample to population benchmarks.

Table 3: Summary Table.

	Mean	Std.Dev.	Min	Max
Woman	0.477	0.500	0	1
College Graduate	0.477	0.500	0	1
18-24	0.231	0.422	0	1
25-34	0.241	0.428	0	1
35-54	0.291	0.454	0	1
+55	0.237	0.425	0	1
Turkish Ethnicity	0.857	0.350	0	1
Minority	0.350	0.477	0	1
Lower SES	0.220	0.414	0	1
Lower-Mid SES	0.305	0.461	0	1
Middle SES	0.417	0.493	0	1
Upper-Mid SES	0.053	0.224	0	1
Upper SES	0.005	0.069	0	1
Previously Voted	0.914	0.281	0	1
Voted for Ruling Coalition	0.420	0.494	0	1
Interested in Politics	0.628	0.484	0	1
Following Political News	0.542	0.498	0	1
Following News on Internet	0.666	0.472	0	1
Feeling Close to Party	0.296	0.457	0	1
Politically Knowledgeable	0.309	0.462	0	1
Full Democratic Support	0.702	0.458	0	1
Partial Democratic Support	0.896	0.305	0	1
Accuracy Feedback Group	0.317	0.465	0	1
Conversation Exposure Group	0.338	0.473	0	1
Placebo Group	0.345	0.476	0	1

Source: Author's Original Survey Experiment.

4 Main Results

4.1 Preregistered Analyses

Primary analyses follow the preregistration and focus on affective polarization and beliefs about out-group extremity as the main outcomes. I then report a limited set of additional analyses examining engagement and heterogeneity by baseline priors, politicization, partisan affiliation, education, and self-reported information sources. Outcomes are standardized using the placebo group mean and standard deviation, so coefficients are reported in placebo-standard-deviation units. I report two-sided p-values and present both 90% and 95% confidence intervals. Appendix 9 reproduces preregistered one-tailed tests; the substantive conclusions are unchanged.

Affective polarization. Figure 5 reports effects on social distance and negative trait associations. For social distance, conversation exposure produces its most favorable social-distance estimate among opposition supporters (-0.126 SD, 90% CI [-0.269, 0.017]), but the interval includes zero. In the full sample, the estimate is also negative and imprecise, while among government supporters it is essentially zero. For negative trait associations, conversation-exposure estimates are imprecise in the full sample and in both partisan subgroups, with no clear evidence of improvement. Descriptively, the social-distance estimate is more negative among opposition supporters than among government supporters, but pooled joint tests do not provide strong evidence of subgroup differences (for joint tests of the treatment-by-incumbent interactions, please see Appendix 10).

Accuracy feedback produces a different pattern. Among government supporters, social distance increases (0.198 SD, 90% CI [0.022, 0.374]), suggestive of a less favorable affective response in this subgroup. For

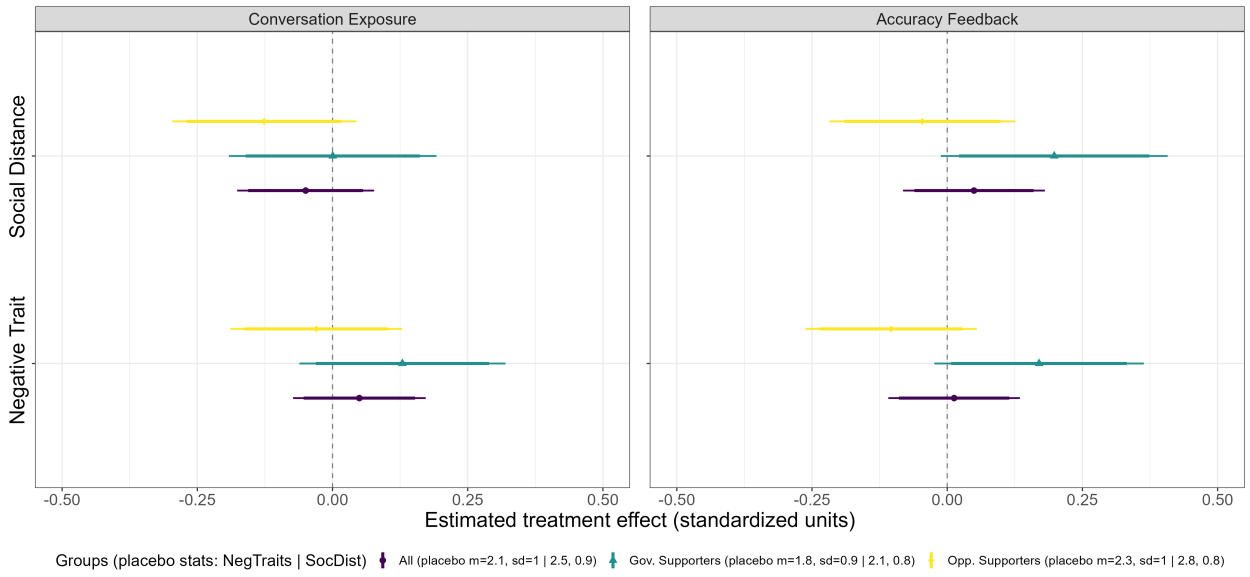

Figure 5: Sample Average Treatment Effects on Social Distance and Negative Trait Associations (N=1263).

opposition supporters and the full sample, the social-distance estimates are not distinguishable from zero at the two-sided 90% level, although the opposition estimate is negative. Negative trait associations show a similar pattern. Among government supporters, the estimate is positive (0.17 SD, 90% CI [0.007, 0.332]), while the opposition and full-sample estimates remain imprecise. Taken together, these results suggest that any adverse affective response to accuracy feedback is concentrated among government supporters rather than shared across partisan subgroups.

Overall, these affective estimates are heterogeneous across partisanship and often imprecise. I next examine perceived out-group extremity, where the interventions were expected to generate more direct updating and where the evidence is correspondingly sharper.

Belief updating about out-group extremity. If the treatments work first by shifting beliefs about out-group extremity, belief outcomes should provide the clearest test of the underlying informational mechanism. Figure 6 reports effects on post-treatment perceived out-group extremity. Accuracy feedback shifts beliefs toward the benchmark distribution in the full sample (-0.103 SD, at 95% CI [-0.194, -0.012]) and among opposition supporters (-0.152 SD, at 95% CI [-0.277, -0.027]). Among government supporters, the point estimate is close to zero and imprecise. Conversation exposure does not produce a clear overall shift in perceived out-group extremity in the full sample (-0.008 SD, at 95% CI [-0.095, 0.08]). Among opposition supporters, however, it shifts beliefs in the corrective direction and the estimate excludes zero at the two-sided 90% level (-0.108 SD, 90% CI [-0.202, -0.014]). By contrast, the estimate for government supporters is positive and imprecise. Point estimates are more consistent with stronger belief correction among opposition supporters than among government supporters, although the pooled interaction test provides only suggestive evidence of subgroup differences. Pooled joint tests of whether the treatment effects differ across partisan subgroups provide suggestive evidence of subgroup differences for the belief outcome at the 10% level (Appendix 10).

Engagement (duration). Figure 7 reports preregistered duration (log time-on-page) for two post-treatment pages. Conversation exposure produces positive full-sample point estimates for time-on-page on both pages, but these estimates are imprecise and their confidence intervals include zero. Accuracy feedback shows no clear changes in time-on-page. Appendix 10 shows no evidence that treatment effects differ across partisan subgroups for the *Extremist Actions* page ($p = 0.612$). For the *Implicit Extremist Sorting* page, the joint test is only marginal ($p = 0.085$), so I treat any subgroup differences as suggestive rather than definitive. Because duration can reflect attention, effort, confusion, or interface friction, I interpret these outcomes as descriptive evidence about engagement with the material rather than as evidence about changes in substantive attitudes.

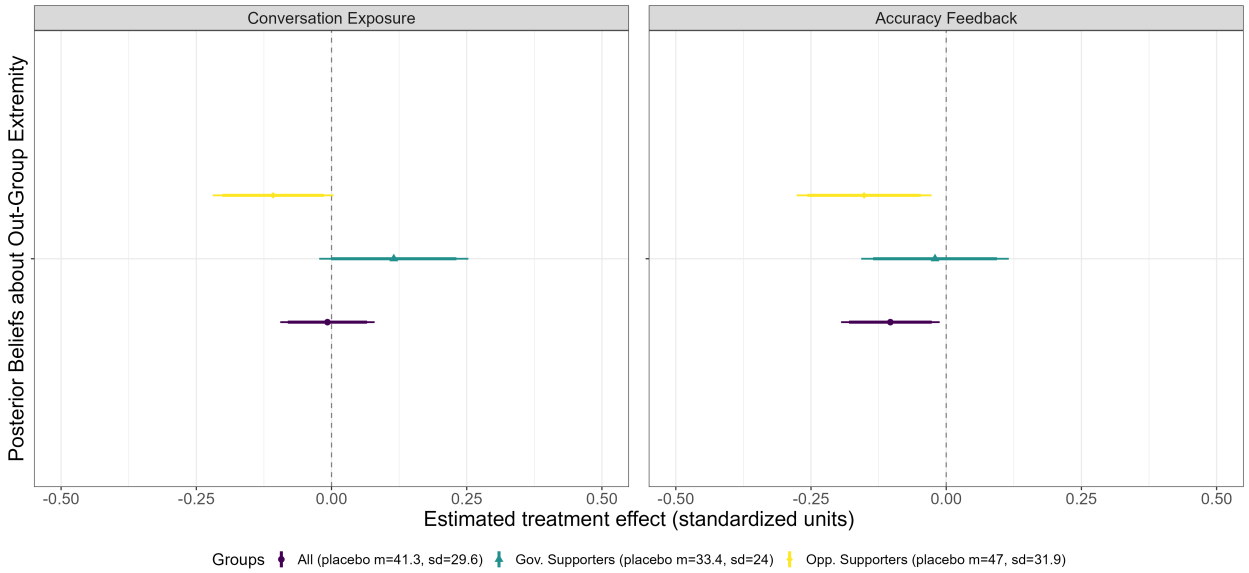

Figure 6: Sample Average Treatment Effects on Posterior Beliefs on Out-Group Extremity. (N=1263).

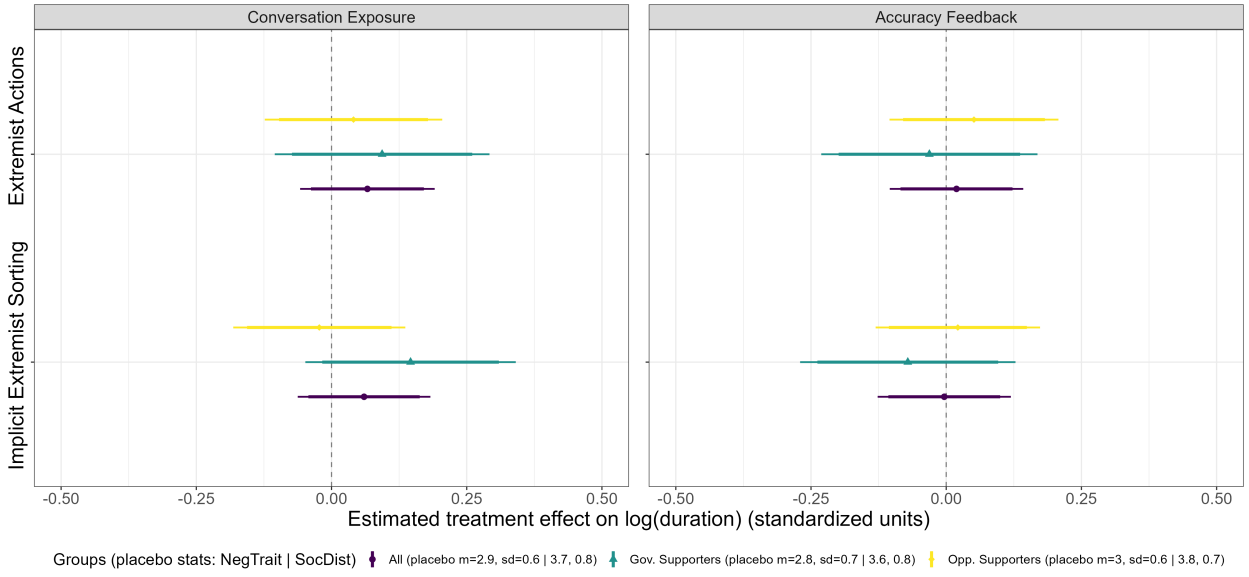

Figure 7: Sample Average Treatment Effects on on Page Duration (log). (N=1263).

Summary. The preregistered results show a clear asymmetry across interventions. Accuracy feedback reduces perceived out-group extremity on the targeted issues, with the strongest corrective estimates appearing among opposition supporters. Conversation exposure produces weaker and more heterogeneous belief updating, with suggestive corrective effects concentrated among opposition supporters rather than the full sample. Affective outcomes do not move uniformly. Conversation exposure produces its most favorable social-distance estimate among opposition supporters, but these estimates are less precise than the corresponding belief effects. Accuracy feedback, by contrast, produces the least favorable affective estimates among government supporters, increasing both social distance and negative trait associations. Duration outcomes are directionally more positive under conversation exposure than under accuracy feedback, but they remain imprecisely estimated and are best interpreted as descriptive evidence about engagement rather than substantive attitudinal change. These subgroup patterns are suggestive of stronger belief correction among opposition supporters, especially for conversation exposure, although pooled subgroup-difference tests remain weak.

4.2 Heterogeneous Treatment Effects

Because the study was not powered to detect modest subgroup differences with precision, I interpret subgroup contrasts below as directional evidence about scope conditions rather than as definitive tests of partisan heterogeneity. I examine heterogeneity in two steps. First, I estimate preregistered moderation models testing whether treatment effects on social distance vary with baseline priors about out-group extremity, politicization, partisan affiliation, and education. Second, motivated by the backsliding framework in which information environments politicize credibility, I examine whether responses differ by respondents' self-reported primary news source. I treat the news source split as descriptive because it was not preregistered.

Table 4: Moderated Linear Regression Results

	Dependent variable:			
	(Social Distance)			
	Prior Belief	Politicization	High Affiliation	Higher Education
	(1)	(2)	(3)	(4)
Accuracy Feedback	0.048 (0.066)	0.137 (0.108)	0.030 (0.078)	0.059 (0.094)
Conversation Exposure	−0.056 (0.064)	0.063 (0.104)	0.002 (0.075)	−0.052 (0.090)
Prior Beliefs	0.006 (0.105)	−0.115 (0.100)	−0.106 (0.100)	−0.114 (0.100)
Squared Prior Beliefs	0.265*** (0.102)	0.261** (0.103)	0.252** (0.103)	0.260** (0.103)
Accuracy x Prior	−0.221*** (0.067)			
Conversation x Prior	−0.158** (0.065)			
Politicization	−0.063 (0.062)	0.035 (0.094)	−0.070 (0.062)	−0.071 (0.062)
Accuracy x Politicization		−0.146 (0.137)		
Conversation x Politicization		−0.182 (0.133)		
High Affiliation	0.412*** (0.064)	0.409*** (0.065)	0.446*** (0.103)	0.408*** (0.065)
Accuracy x Affiliation			0.065 (0.151)	
Conversation x Affiliation			−0.169 (0.147)	
Higher Education				−0.011 (0.091)
Accuracy x Education				−0.019 (0.134)
Conversation x Education				0.005 (0.130)
Observations	1,263	1,263	1,263	1,263
Adjusted R^2	0.111	0.104	0.105	0.102

Two-sided tests. (N = 1263).

Table 4 reports preregistered moderation models using two-sided tests. Baseline prior beliefs are the only preregistered moderator that shows clear evidence of conditioning treatment effects on social distance. The interaction terms are negative for both accuracy feedback and conversation exposure, implying that both interventions become more socially distance-reducing as prior extremity beliefs increase. The evidence is stronger for accuracy feedback than for conversation exposure, but both interactions exclude zero at the 95%

level. I interpret this pattern as substantively meaningful but not uniformly general, because the subgroup plots reported below suggest that it is driven mainly by opposition supporters and is estimated imprecisely among government supporters. By contrast, the evidence for moderation by politicization, partisan affiliation strength, and education is weak. Their interaction estimates are imprecise, include zero, and do not provide clear evidence that treatment effects differ systematically across these respondent characteristics.

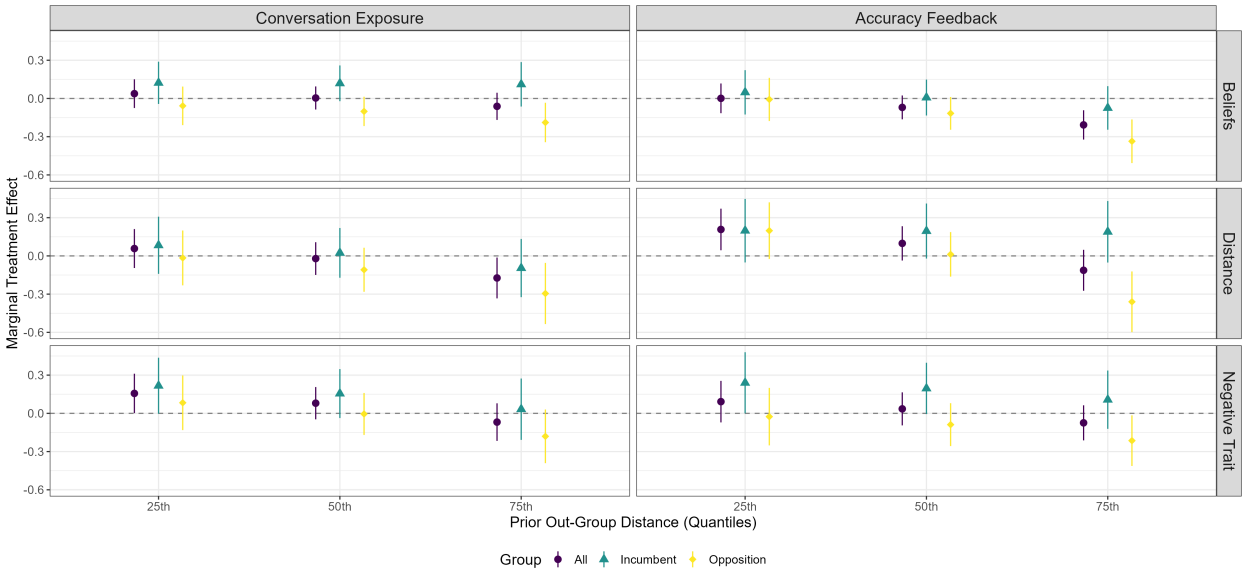

Figure 8: Marginal SATE on Posterior Beliefs, Social Distance, and Negative Trait Association (standardized) at Prior Beliefs at 25th, 50th, and 75th Percentiles. (N=1263).

Figure 8 complements the regression models by plotting marginal treatment effects at the 25th, 50th, and 75th percentiles of baseline priors, computed separately within partisan subgroups. Moving from the 25th to the 75th percentile means moving from relatively low to relatively high prior extremity. The main pattern is not a universal monotonic gradient, but a subgroup-specific one. Among opposition supporters, both interventions become more favorable as prior extremity rises. This is clearest for social distance: accuracy feedback shifts from positive at low prior levels to more negative at the 75th percentile, where the estimate excludes zero at the 95% level, and conversation exposure also reduces social distance more at higher priors. Belief updating follows a similar logic, with the strongest corrective effects also appearing among opposition supporters at the upper end of the prior distribution. By contrast, government supporters do not show the same degree of social distance decrease, and the full-sample pattern is weaker. Negative trait associations are less coherent across percentiles and groups, so I treat those plots as descriptive rather than as evidence of a stable moderation pattern. Overall, the quartile analysis reinforces the idea that prior-belief moderation is clearest for beliefs and social distance, and that it is driven mainly by opposition supporters.

To complement the coefficient-by-coefficient results in Table 4 and Figure 8, Appendix 10 shows joint tests of the two treatment-by-moderator interactions for each preregistered moderator. Only baseline prior beliefs show clear evidence of moderation ($p = 0.003$). Both interaction coefficients are negative, consistent with Figure 8, where treatment effects become more social-distance-reducing as prior beliefs about out-group extremity increase. By contrast, the joint tests for politicization, partisan affiliation strength, and higher education are not statistically distinguishable from zero, so I do not interpret those moderators as substantively meaningful in this study.

Finally, I conduct an exploratory, non-preregistered subgroup analysis by self-reported mainstream media use within each partisan subgroup (Figure 9). The descriptive patterns are not uniform, but they are suggestive. Among opposition supporters who report mainstream news, both interventions produce more corrective belief estimates than among opposition supporters who do not, and the corresponding social-distance estimates are also more favorable. Among government supporters who report mainstream news, the clearest adverse pattern appears on affective outcomes: negative trait associations increase under both interventions,

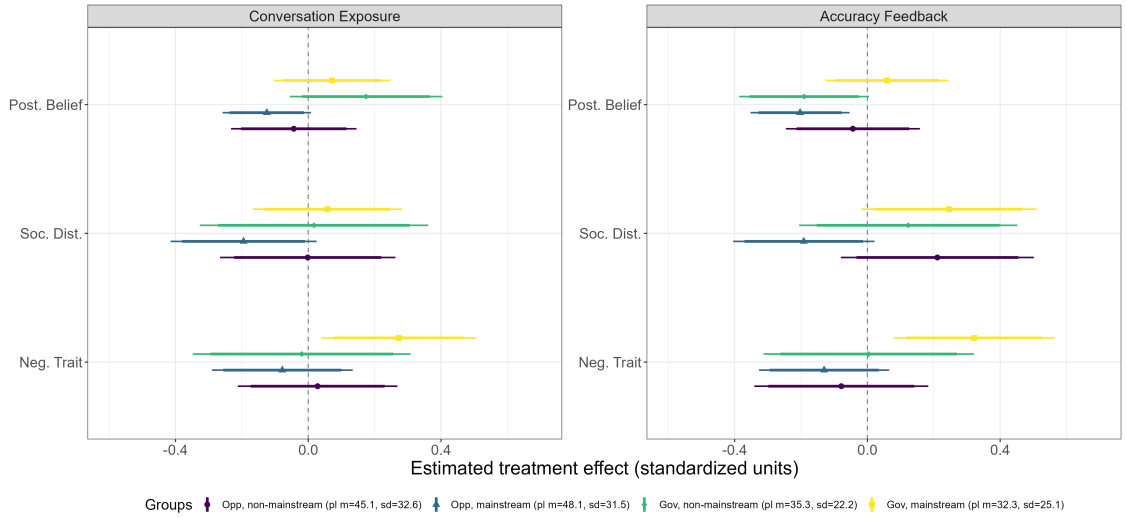

Figure 9: Sample Average Treatment Effects. (N=1263).

and social distance also increases under accuracy feedback. I do not interpret this as confirmatory evidence that mainstream media exposure itself causes these differences as self-reported mainstream news use is a coarse and non-randomized measure. The narrower interpretation I can defend is that treatment responses vary across politically differentiated information environments, in ways that are consistent with the paper's broader argument about credibility and threat.

5 Discussion

Can stereotype-challenging information reduce hostility in a backsliding democracy? This study argues that the answer is more nuanced than either optimism or pessimism would allow. Targeted, issue-specific corrections can alter beliefs about the political out-group, yet those belief changes do not reliably yield warmer intergroup attitudes. In a preregistered survey experiment conducted in Türkiye, accuracy feedback led respondents in the full sample to see the political out-group as less extreme, whereas conversation exposure produced weaker average belief updating. The affective outcomes are less settled. Those estimates are more heterogeneous and, in many cases, less precise. What emerges, then, is that corrective information can reshape political perceptions even in a highly polarized setting, while its effects on hostility remain less secure and more dependent on both audience and form.

Stereotype correction is not confined to consolidated democracies with relatively trusted information environments. Türkiye offers a demanding setting in which to evaluate such interventions. Media are politicized, credibility is contested, and citizens have ample reason to approach corrective information defensively. If a brief intervention can nonetheless shift perceived out-group extremity in such a setting, that is substantively important. The study also underscores the value of targeting. Rather than asking respondents to revise their views of the out-group in the abstract, the interventions focus on specific policy domains in which exaggeration is especially likely. The fact that brief corrections on those issues shift broader perceptions of out-group extremity suggests that the selected domains were consequential rather than incidental. This does not establish that the RDPD procedure is uniquely optimal. However, the procedure can identify issues on which stereotype correction has meaningful leverage.

The results also make clear that belief correction and affective depolarization should not be treated as interchangeable outcomes. Accuracy feedback is the stronger instrument for moving beliefs because it directly supplies base-rate information. Yet for that same reason, it may also be the more affectively risky format in an environment where credibility is itself politicized. Conversation exposure works differently. It conveys the same underlying signal of out-group moderation in a more socially familiar and less overtly corrective form. In this study, that format appears weaker on average as a vehicle for factual updating, yet in some subgroup

analyses it yields more favorable relational estimates. Format, in other words, is not merely a matter of presentation. It helps shape whether corrective information is received as evidence, as threat, or as some unstable mixture of the two.

While I do not treat the subgroup differences as definitive because the study was not powered to estimate every subgroup contrast precisely, the directional pattern is suggestive. Government supporters show the least favorable affective estimates, particularly under accuracy feedback. In the main results, neither intervention yields clear evidence of affective improvement in this subgroup, and accuracy feedback is associated with increases in both social distance and negative trait associations. That pattern matters because it suggests that stereotype correction does not fail only by leaving attitudes unchanged. Under some conditions, it may also provoke a more negative response, even when the information itself points toward moderation. I do not take this to mean that government supporters are simply unreachable. The quartile results suggest a narrower conclusion: receptivity is shaped unevenly by prior beliefs, and the strongest reductions in social distance appear among opposition supporters, who also begin from more distorted priors on average, rather than among government supporters.

The moderation results deepen that point by complicating a familiar pessimism, namely the view that citizens with the most distorted priors must also be the least responsive to correction. In this study, baseline priors are the only preregistered moderator that clearly moderates treatment effects on social distance. Yet, the resulting pattern does not display the clean attenuation at high prior extremity like the preregistered non-monotonic expectation. If anything, more favorable treatment effects often appear at higher levels of prior extremity, especially within the opposition subgroup. Findings indicate that the scope for updating under backsliding may remain politically meaningful precisely where distortions are greatest. The larger lesson is not that stronger prior misperceptions mechanically produces more change, but that prior beliefs matter through audience-specific pathways shaped by partisan identity, perceived credibility, and intervention format.

The affective outcomes themselves do not move as a single block. Social distance appears more malleable overall, especially among opposition supporters and at higher levels of prior extremity. Negative trait associations are less responsive in the aggregate, but they are not uniformly flat across subgroups. Most notably, accuracy feedback increases negative trait associations among government supporters, suggesting that adverse affective responses in that subgroup are not confined to social distance alone. This, too, is clarifying. Affective polarization is not a single attitudinal object, and different intervention formats may shift its relational and moralized components in different ways across audiences.

The exploratory media analysis points in a similar direction, though it should be read with caution. In the standard interaction models, treatment differences by mainstream news use are not estimated precisely. In the descriptive subgroup analyses, however, the pattern is more textured than a simple backlash account would imply. Among opposition supporters who report mainstream news use, both interventions produce more corrective belief estimates than among opposition supporters who do not, and the corresponding reduction in social distance is also larger. Among government supporters who report mainstream news use, the clearest adverse pattern appears on the affective outcomes: negative trait associations rise under both interventions, and social distance increases as well under accuracy feedback. I do not interpret this as confirmatory evidence of media causality. Self-reported primary news source is a crude measure, and the analysis was not preregistered. The narrower claim I can defend is that treatment responses vary across politically differentiated information environments.

How, then, should the backlash pattern among government supporters be understood? One plausible interpretation is a form of double updating. Respondents may revise beliefs about the empirical world while also revising beliefs about communicator intent. In a politicized credibility environment, numerical corrections may be processed not only as information, but also as social signals about who is entitled to judge, correct, or audit the in-group. In Rozenas and Stukal's framework, exposure to conflicting information can induce "double updating," whereby recipients update both on the state of the world and on the reliability or motives of the communicator (Rozenas and Stukal 2019). In a backsliding context, that mechanism could generate precisely the pattern observed here: more accurate beliefs without warmer affect, and perhaps even greater distance if the correction is experienced as hostile scrutiny. A related possibility is that propaganda works less by persuading citizens of every factual claim than by stabilizing the interpretive frames through which counter-attitudinal information is received (Shirikov 2024). My design does not randomize messenger identity, endorsement cues, or credibility labels, so these mechanisms remain suggestive rather than identified. The

empirical point I can defend more directly is narrower, but no less important: factual updating and affective response do not move in lockstep, and among regime-aligned respondents the relationship between the two can be negative.

Taken together, the findings point to a genuine tradeoff for depolarization under democratic backsliding. Accuracy feedback is the stronger corrective tool for beliefs about out-group extremity, but it is not affect-safe for all audiences, and among government supporters it coincides with adverse movement in both social distance and negative trait associations. Conversation exposure is a weaker corrective tool for beliefs on average, but it produces its most favorable relational estimates among opposition supporters, particularly among those with more extreme baseline priors. More broadly, the results suggest that stereotype correction is not a one-dimensional intervention problem. Its effects depend on who receives the message, how the message is delivered, and the credibility environment in which it is embedded.

This paper also has several limitations. It measures attitudes shortly after exposure and therefore cannot establish persistence or downstream behavioral consequences. Subgroup analyses and three-way splits are informative, but they reduce precision, and the design was not powered to detect every modest heterogeneous effect with confidence. The news-source analysis is exploratory and relies on coarse self-reported categories that compress substantial variation in exposure and trust. Duration is, by its nature, an ambiguous proxy for engagement. Finally, the design does not vary messenger identity, endorsement cues, or credibility labels, all of which would help adjudicate the mechanisms more cleanly.

These limitations do not diminish the paper's central contribution. The study provides evidence from a backsliding democracy that targeted corrections on stereotype-prone issues can shift perceived out-group extremity, even when affective polarization does not reliably recede as beliefs become more accurate. More broadly, it shows that in such settings targeted correction can improve beliefs about the political out-group, but whether those belief changes translate into lower hostility depends on intervention format and audience-specific scope conditions. In polarized environments with state-captured media, the central question is not simply whether citizens can be corrected. It is whether a given corrective format can produce factual updating without simultaneously activating identity threat. Future work should build on this by randomizing messenger identity and credibility cues, tracing durability under repeated exposure, and measuring perceived threat and trust at the moment of updating. For now, the evidence supports a disciplined conclusion: correcting stereotypes about out-group extremity is feasible even under democratic erosion, but reducing hostility requires designs that treat source credibility, audience psychology, and prior beliefs as first-order constraints rather than as secondary implementation details.

6 References

Ahler, Douglas J, and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80 (3): 964–81.

Barberá, Pablo. 2020. "Social Media, Echo Chambers, and Political Polarization." *Social Media and Democracy: The State of the Field, Prospects for Reform* 34.

Berger, Ronald J, and Richard Quinney. 2004. *Storytelling Sociology: Narrative as Social Inquiry*. Lynne Rienner Publishers.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics* 131 (4): 1753–94.

———. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110 (10): 2997–3029.

Chaudoin, Stephen, Brian J Gaines, and Avital Livny. 2021. "Survey Design, Order Effects, and Causal Mediation Analysis." *The Journal of Politics* 83 (4): 1851–56.

Druckman, James N, Suji Kang, James Chu, Michael N. Stagnaro, Jan G Voelkel, Joseph S Mernyk, Sophia L Pink, Chrystal Redekopp, David G Rand, and Robb Willer. 2023. "Correcting Misperceptions of Out-Partisans Decreases American Legislators' Support for Undemocratic Practices." *Proceedings of the National Academy of Sciences* 120 (23): e2301836120.

Druckman, James N, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2022. "(Mis) Estimating Affective Polarization." *The Journal of Politics* 84 (2): 1106–17.

Erdoğan, Emre, and Nora Fisher Onar. 2020. "Dimensions of Polarization in Turkey 2020." *The German Marshall Fund of the United States*.

Esen, Berk. 2021. "Competitive Authoritarianism in Turkey Under the AKP Rule." In *The Routledge Handbook on Contemporary Turkey*, 153–67. Routledge.

Gessler, Theresa, and Natasha Wunsch. 2023. "A New Regime Divide? Partisan Affect and Attitudes Towards Democratic Backsliding."

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.

Hartman, Rachel, Will Blakey, Jake Womick, Chris Bail, Eli J Finkel, Hahrie Han, John Sarrouf, et al. 2022. "Interventions to Reduce Partisan Animosity." *Nature Human Behaviour* 6 (9): 1194–1205.

İmamoğlu, Umut Koray. 2022. "To Vote or Not to Vote? Affective Polarization and Voter Turnout in Turkey." PhD thesis.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–31.

Iyengar, Shanto, and Sean J Westwood. 2015. "Fear and Loathing Across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59 (3): 690–707.

Jenke, Libby. 2023. "Affective Polarization and Misinformation Belief." *Political Behavior*, 1–60.

Kahan, Dan M. 2013. "Ideology, Motivated Reasoning, and Cognitive Reflection." *Judgment and Decision Making* 8 (4): 407–24.

Klar, Samara, Yanna Krupnikov, and John Barry Ryan. 2018. "Affective Polarization or Partisan Disdain? Untangling a Dislike for the Opposing Party from a Dislike of Partisanship." *Public Opinion Quarterly* 82 (2): 379–90.

Kutlay, Mustafa, and Ziya Öniş. 2024. "Governance Crises and Resilience of Authoritarian Populism: 2023 Turkish Elections from the Perspective of Hirschman's 'Exit, Voice, and Loyalty'." *Southeast European and Black Sea Studies*, 1–21.

Leemann, Lucas, Lukas F Stoetzer, and Richard Traunmüller. 2021. "Eliciting Beliefs as Distributions in Online Surveys." *Political Analysis* 29 (4): 541–53.

Lees, Jeffrey, and Mina Cikara. 2020. "Inaccurate Group Meta-Perceptions Drive Negative Out-Group Attributions in Competitive Contexts." *Nature Human Behaviour* 4 (3): 279–86.

Maniadis, Zacharias, Fabio Tufano, and John A List. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104 (1): 277–90.

McCoy, Jennifer, Tahmina Rahman, and Murat Somer. 2018. "Polarization and the Global Crisis of Democ-

racy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities." *American Behavioral Scientist* 62 (1): 16–42.

Moral, Mert. 2017. "The Bipolar Voter: On the Effects of Political Polarization on Voter Turnout and Voting Behavior." PhD thesis, State University of New York at Binghamton.

Neundorf, Anja, and Aykut Öztürk. 2021. "Recruiting Research Participants Through Facebook: Assessing Facebook Advertisement Tools."

Nyhan, Brendan. 2021. "Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions." *Proceedings of the National Academy of Sciences* 118 (15): e1912440117.

Orhan, Yunus Emre. 2022. "The Relationship Between Affective Polarization and Democratic Backsliding: Comparative Evidence." *Democratization* 29 (4): 714–35.

Rozenas, Arturas, and Denis Stukal. 2019. "How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television." *The Journal of Politics* 81 (3): 982–96.

Schotter, Andrew, and Isabel Trevino. 2014. "Belief Elicitation in the Laboratory." *Annu. Rev. Econ.* 6 (1): 103–28.

Shiller, Robert J. 2017. "Narrative Economics." *American Economic Review* 107 (4): 967–1004.

Shirikov, Anton. 2024. "Rethinking Propaganda: How State Media Build Trust Through Belief Affirmation." *The Journal of Politics* 86 (4): 1319–32.

Somer, Murat. 2019. "Turkey: The Slippery Slope from Reformist to Revolutionary Polarization and Democratic Breakdown." *The ANNALS of the American Academy of Political and Social Science* 681 (1): 42–61.

Svolik, Milan W. 2019. "Polarization Versus Democracy." *Journal of Democracy* 30 (3): 20–32.

Taber, Charles S, and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–69.

Tajfel, Henri, John C Turner, William G Austin, and Stephen Worchel. 1979. "An Integrative Theory of Intergroup Conflict." *Organizational Identity: A Reader* 56 (65): 9780203505984–16.

Turner, John C. 1975. "Social Comparison and Social Identity: Some Prospects for Intergroup Behaviour." *European Journal of Social Psychology* 5 (1): 1–34.

Vegetti, Federico, and Moreno Mancosu. 2020. "The Impact of Political Sophistication and Motivated Reasoning on Misinformation." *Political Communication* 37 (5): 678–95.

Vives, Marc-Lluís, Mina Cikara, and Oriel FeldmanHall. 2022. "Following Your Group or Your Morals? The in-Group Promotes Immoral Behavior While the Out-Group Buffers Against It." *Social Psychological and Personality Science* 13 (1): 139–49.

Voelkel, Jan G, Michael Stagnaro, James Chu, Sophia Pink, Joseph Mernyk, Chrystal Redekopp, Isaias Ghezae, et al. 2023. "Megastudy Identifying Effective Interventions to Strengthen Americans' Democratic Attitudes."