From Correction to Connection: Relational Approaches to Countering Misinformation*

Cesi Cruz Julien Labonne

July 2026

Abstract

Most interventions against misinformation are designed to improve individual accuracy: correcting false beliefs or training people to evaluate claims more carefully. We argue that misinformation is a feature of relational information environments, in which political information is socially mediated. In a randomized field experiment in 160 villages in the Philippines, we test workshops that bring communities together to reflect collectively on information challenges, discuss shared practices for sharing information, and navigate disagreement. Individuals in treated villages shifted their sources of political news, relying less on social media and engaged more frequently in offline discussion through civic and religious networks. In addition, individuals in treatment areas perceived less political polarization and reported greater interest in politics. These effects extended through social networks, reaching people who are connected to treated individuals even when they did not attend the workshops themselves. Effects also extend to electoral outcomes: official electoral results indicate that candidates who relied heavily on social media advertising underperformed in treatment villages, and voters in those areas were less responsive to false online rumors about local candidates. These results suggest that misinformation is not only an informational problem but a social one, necessitating interventions to support communities, not only individuals.

*Cruz: University of Michigan (cesi@umich.edu). Labonne: University of Oxford and Centre for Economic Policy Research (julien.labonne@bsg.ox.ac.uk). We are grateful to Jerold Asuncion, Rachel Berwald, Jose Gloria, Dan Harker, Judy Malabanan, and Nikita Savin for excellent research assistance and to Prudenciano Gordoncillo, Charis Tolentino and the UPLB team for collecting the data. We thank Nikola Banovic, Alain Cohn, Licheng Liu, Franshelly Martinez-Ortiz, and Francesco Trebbi for comments. We also benefited from feedback from seminars at UCLA, UC Berkeley, the University of Michigan, and the University of Oxford. The project received ethics approval from Oxford and UCLA.

1 Introduction

People do not encounter political information in a vacuum. It is filtered and interpreted through social relationships that govern what sources are credible and what narratives persist. Yet most efforts to counter misinformation and polarization focus on individual discernment: training citizens to distinguish truth and falsehood.

The record of this individual approach is mixed, against a backdrop of developments in social media and generative AI that exacerbate the challenges by lowering the cost to produce and disseminate false content. Fact-checks, labels, accuracy prompts, and media literacy training can improve discernment in the short run (Guess et al., 2020; Pennycook and Rand, 2019; Porter and Wood, 2021). However, some of these interventions have effects that fall along partisan lines, and in some settings even backfire, hardening beliefs even further. (Nyhan and Reifler, 2010; Bolsen and Druckman, 2018; Gottlieb et al., 2022; Jenke, 2024). Nearly all of the evidence also comes from Western democracies, where trust in mainstream media and regulators is comparatively high (Blair et al., 2024).

We start from a different premise: misinformation is not only a failure of individual cognition, but a feature of *relational* information environments. Political information circulates through families, churches, friendship networks, and community groups, where messages carry social meaning beyond their factual content (Mason, 2018; Cinelli et al., 2021; Mosleh et al., 2021). Trust in information is inseparable from trust in the messenger, and sharing information often serves relational functions, such as signaling care or asserting group belonging.

As a result, engagement with misleading content also increases its visibility and circulation within the community, making these choices about information interdependent. This implies that misinformation persists not only because individuals struggle to assess the accuracy of the content they encounter, but because community incentives shape how political information is shared and discussed.

This perspective also suggests a different unit of intervention. Rather than treating individuals as isolated consumers of information, we treat communities as the sites where norms about information sharing, discussion, and trust are formed and reinforced. The goal is not to produce better fact-checkers, but to strengthen the relational contexts that shape how political information is gathered, discussed, and acted upon.

We evaluate this account with a randomized field experiment in 160 villages in the Philippines, one of the earliest and most visible cases of social-media driven misinformation (Tapsell, 2021; Arugay and Mendoza, 2024). Villages were randomly assigned

to receive an in-person workshop that brought communities together to discuss and reflect collectively on their information environment. Using media literacy as a starting point for discussion, participants were encouraged to reflect on their community's practices and norms around sharing information and navigating disagreement. We surveyed respondents at baseline, about one month after the workshops, and about nine months after, together with a separate sample of people whom baseline respondents had named as regular discussion partners, and we link villages to official precinct-level returns from the May 2025 elections.

Treated respondents perceived the country as less divided and reported more interest in politics. They also relied less on social media for political news: nine months out, they were 14 percentage points less likely to use YouTube, 8 points less likely to use TikTok, and 4 points less likely to use Facebook, and these effects persisted through endline. They talked more about the news with members of their church and social groups and less with village officials.

These effects produced spillovers throughout the community, reaching people who did not participate in the workshops directly. The discussion partners that baseline respondents named before treatment—none of whom attended a workshop—also used YouTube and TikTok less for news and were more likely to discuss it in church and social groups.

Measuring discernment across different kinds of claims shows gains on some and losses on others. Treated respondents were more resistant to familiar, ideologically-coded narratives and reasoned more accurately about well-established facts like climate change — but they were no better, and if anything worse, at evaluating claims actively circulating on platforms at the time of the survey, a trade-off consistent with reduced exposure to whatever happened to be circulating.

Importantly, these behavioral and attitudinal shifts translated into measurable electoral outcomes. Using official precinct-level results from the 2025 elections, about a year after the workshops, we find that candidates who relied heavily on Facebook advertising underperformed in treatment villages. Treatment villages were also less responsive to false online rumors about local candidates. These findings suggest that changing how communities engage with political information can weaken the effectiveness of digital campaigning and disinformation strategies.

Our work contributes to the research on misinformation in three main ways. First, we bring together two literatures that have largely developed apart: work on political networks, which has long treated political information as socially mediated, flowing through personal influence and discussion networks (Katz and Lazarsfeld, 1955;

Huckfeldt and Sprague, 1995), and work in political psychology, which has advanced theories of individual cognition (Kunda, 1990; Taber and Lodge, 2006; Flynn et al., 2017). We show that interventions designed to reshape how communities engage with political information—where people get news and with whom they discuss it—can generate durable behavioral change and meaningful spillovers (Allcott et al., 2020; Levy, 2021).

Second, we extend the evidence base beyond the Global North. While recent work has begun to examine media literacy and prebunking interventions in the Global South (e.g., Amar et al., 2025; Parihar et al., 2025; Badrinathan, 2021; Dugas et al., 2024; Bowles et al., 2025), much of the evidence on misinformation interventions comes from media environments where trust in mainstream outlets and regulators is relatively high. We study the Philippines, a context characterized by heavy social media use, weak content moderation, and limited media oversight, and show that a community-based intervention can produce durable behavioral change and downstream electoral effects.

Third, we leverage a field experiment and a panel survey tracking respondents over nine months with official precinct-level election returns, allowing us to observe: (i) whether the intervention changed political behavior over time; and (ii) whether these changes affected electoral results .

Taken together, our results indicate that the community is a promising unit of intervention for building resilience to misinformation. The workshops changed how communities accessed and discussed political information, with changes that extended through social networks to downstream effects on vote shares nearly a year later.

2 Misinformation as a Relational Problem

Most existing interventions to counter misinformation conceptualize the problem at the level of individual cognition. In this framework, susceptibility to false or misleading content reflects limitations in individuals' ability to evaluate information accurately, whether due to lack of knowledge, inattention, or cognitive biases. Accordingly, the dominant policy responses, such as fact-checking, content labeling, accuracy prompts, and media literacy training, are designed to improve individual discernment and belief accuracy.

A large literature evaluates the effectiveness of such interventions in reducing belief in false claims or improving the ability to distinguish true from false information. These approaches have generated important insights and, in some contexts, measurable

short-run improvements in belief accuracy and discernment (Lazer et al., 2018; Porter and Wood, 2021; Guess et al., 2020; Pennycook and Rand, 2019).

At the same time, a growing body of work highlights the limits of individual-level corrective approaches. Someinterventions focused on fact checking or belief correction often produce mixed or short-lived effects, particularly in polarized political contexts (Bolsen and Druckman, 2018; Jenke, 2024; Gottlieb et al., 2022). Corrections can also have differential effects based on existing priors, with smaller effects for partisan-based misinformation than for those who are more generally uninformed (Li and Wagner, 2020).

Research on digital media further underscores the importance of information environments rather than individual exposure to misinformation (Gentzkow and Shapiro, 2011). Digital platforms shape not only what information is available, but how it circulates through social networks. These online information flows are often characterized by ideological segmentation and homophily, reflecting both user preferences and network structure (Cinelli et al., 2021; Mosleh et al., 2021). Selective consumption and targeted content reinforce these patterns, incentivizing tailored content to increasingly partisan audiences (Peterson and Kagalwala, 2021; Chopra et al., 2024; Kobayashi et al., 2024; Yarchi et al., 2021). Algorithmic bias and social media platform incentives also privilege engagement over accuracy, increasing the visibility of content that people will react to and share(Vosoughi et al., 2018; Brady et al., 2017; Levy, 2021).

2.1 A community-based framework for misinformation resilience

This literature suggests that misinformation is not only a feature of individual cognition, but a feature of the information environment itself. Families, churches, and civic groups are where norms about sharing, discussing, and trusting information are formed (Banerjee et al., 2013, 2018). Interventions that target individual belief accuracy alone are not able to address the broader social incentives that govern sharing and discussing information.

This paper extends this literature by evaluating a community-based intervention designed to reshape how political information is encountered and discussed within local networks. Rather than attempting to correct specific false beliefs in individuals, the intervention also creates space for discussion and reflection on information sharing practices.

The relational approach we use has an additional testable implication: if resilience to misinformation operates through shared norms rather than individual learning,

treating a subset of a community's residents should also change the behavior of untreated peers. Observing such spillovers would point to a shift in the local information environment, not isolated individual learning.

3 Context, Experiment and Data

3.1 Misinformation in the Philippines

Online disinformation poses a growing threat to democratic governance, particularly in less developed democracies where regulatory and institutional safeguards are weak (Larreguy and Raffler, 2025). Despite the higher stakes in these countries, the evidence base on misinformation interventions is largely limited to the Global North, where institutions, information environments, and media literacy levels differ substantially.

In this context, the Philippines provides an important test case. Often referred to as "Patient Zero" in the global disinformation epidemic, the country has experienced a dramatic shift in political information environments driven by the rapid rise of social media (Tapsell, 2021; Arugay and Mendoza, 2024). Platforms such as Facebook, YouTube, and TikTok are increasingly the primary sources of political news, especially in areas where trust in traditional media is weak. Coordinated disinformation campaigns frequently use local accounts to promote narratives tocast doubt on mainstream media or official government accounts (Arugay and Mendoza, 2024). In early 2025, 67 percent of Filipino respondents reported concern about online misinformation and disinformation. (Chua, 2025b,a).

Election-related misinformation in the Philippines has further intensified in recent years, especially during the 2022 presidential and 2025 midterm elections. Misinformation is also exacerbated by the Philippines' unusual information environment: a population that is both highly digitally connected, but also increasingly distrustful of mainstream information sources (Tapsell, 2021; Ong and Tapsell, 2022). Filipinos find it difficult to tell which news stories are true and which are not (Philippine News Agency, 2025). Leading up to the 2025 elections, misleading content spread across both social media and traditional news outlets. Coordinated influence operations used networks of fake accounts and automated posts to change perceptions of key political candidates and institutions (Arugay and Mendoza, 2024).

These dynamics can also be seen at the community level. In several areas, baseless rumors—such as false claims of candidate disqualification or criminal charges—

circulated widely online and fostered cynicism toward democratic institutions.(Tapsell, 2021; Arugay and Baquisal, 2022; Ong, 2022; Berlinski et al., 2023)

3.2 The experiment

The field experiment was implemented in January to March 2024. The primary objective was to foster community-level norms around how political information is shared and discussed, using media literacy training as a vehicle for starting those conversations (Wardle and Derakhshan, 2017; Guess et al., 2020; Badrinathan, 2021). Rather than training individuals to become better fact-checkers in isolation, the workshops created shared space for communities to reflect collectively on their information environment.

To ground these discussions, we introduced six common techniques used to spread misinformation that were particularly relevant to the Philippine context: (i) Impersonation, (ii) Emotional Language, (iii) Polarization, (iv) Conspiracy Theories, (v) Gaslighting and, (vi) Trolling. To make the concepts more memorable, we designed mascots for each technique. Importantly, the presence of these techniques in a post does not imply that it is necessarily false or misleading—rather, they served as a shared vocabulary for participants to use when discussing and evaluating information within their communities.

The basic treatment is a two hour in-person workshop in groups of about 10, delivered by graduate students and academics at the University of the Philippines in Los Banos. Each workshop started with a small ice-breaking activity, followed by a short introduction to misinformation and by a lecture-style introduction to the concepts. Participants were then provided a number of recent social media posts and news stories and were asked to discuss in groups whether the posts used some of the common techniques and whether they believed the stories or not.

A number of design choices are important to discuss. First, we opted for in-person rather than online interventions as we thought that it would give participants more time to process the material and ask clarifying questions.[1] To reduce concerns about low take-up the field team selected dates to be convenient for participants. In addition, they were offered a small compensation for their time and the team sent regular reminders in the weeks leading up to the sessions. In the end, 95.5% of invited participants joined the workshops.

Second, the workshops balanced instruction with interactive time for participants to

[1] We considered having 3 to 4 sessions spread over a few weeks but it was too challenging to coordinate participant schedules and so we settled for one session.

engage with each other and practice their newly acquired skills with example posts. This was meant to foster discussions, increase retention and increase the likelihood that participants would put the information into practice in their day-to-day conversations with their broader social networks. The example posts used did not include political content. In practice, about 30 minutes were spent on instruction with the remainder used for discussion and practice.

Third, we opted for small groups to create space for individuals to discuss their concerns about dis/misinformation. We were concerned that participants would be less likely to bring up their concerns in larger groups. Participants were provided with a key takeaway document that they could bring home and were encouraged to discuss those issues with their friends and relatives.

Fourth, the intervention timing, about half way between the May 2022 and May 2025 national elections, was meant to coincide with a time when issues of misinformation online were not as salient and politicized.

A subset of treated villages received add-ons. The second treatment arm added a small activity at the end: participants were taught healthy/non-confrontational ways to approach someone they suspect may be spreading misinformation. The third treatment arm started with a discussion on important Filipino values and of whether misinformation has played a role in harming these common values. These variants are intended to test the sensitivity of the treatment to the content of the training modules.

Operationally, the sample of 160 villages was divided, within municipalities, in groups of four based on vote share received by President Bongbong Marcos in the 2022 elections.[2] In each quadruplet, villages were randomly allocated to control and to one of the 3 treatment groups. In each treatment village 10 individuals were invited to participate. The groups were well-balanced.

3.3 Data

Our analysis relies on four main sources of data. We implemented three individual-level surveys and collected official precinct-level results for the May 2025 elections from COMELEC.

The baseline survey was implemented in September to October 2023 with a sample of 1,600 respondents across 160 villages. All baseline respondents in our treatment villages were invited to the workshops. Respondents were sampled from official

[2]List of municipalities is available in Table A.1

voters lists in each village. The data include information on access to a number of news sources and whether they are used to get information about entertainment, news and politics. Options include both mainstream media, social media platforms, and friends/relatives. Respondents were asked to rate a number of specific news sources for trustworthiness and entertainment values. We also collected data on respondents' emotional state. Finally, respondents were asked to rate the trustworthiness of and to indicate whether they believed a number of metanarratives that were circulating online at the time of the survey.

The midline survey was implemented between March to May 2024 and the endline survey in December 2024-January 2025. The team managed to re-interview 98.5% of baseline respondents at midline and 97.7% of baseline respondents at endline. Attrition and its determinants are stable across treatment and control groups. In addition to the variables measured at baseline, we also collected data on beliefs about climate change

In addition, to estimate spillover effects, baseline respondents were asked to provide the name and contact details of one individual with whom they talked regularly and who lived in the same village. Those individuals were interviewed at endline, and the process used to identify and reach out to them was identical in treated and control villages.[3] Due to budget constraints, we could only collect data on 1,200 spillover respondents and on a subset of variables and made sure to include measures of news consumption.

Finally, we also use precinct-level results from the May 2025 elections downloaded from the official COMELEC website.[4] Precincts are nested within villages and there is at least one precinct per village. We can thus perfectly allocate precincts to treatment and control villages.

A timeline of the experiment and associated surveys is available in Figure 1.

4 Main results

We start by evaluating the main effects of our workshops by estimating equations of the form:

$$Y_{ivqt} = \beta^m T^m_{vqt} + \beta^e T^e_{vqt} + \eta_t + \omega_q + \epsilon_{ivqt}$$

[3]None of individuals identified through this process were other baseline respondents.
[4]https://2025electionresults.comelec.gov.ph/dashboard

where Y_{ivq} is the outcome of interest for individual i in village v in quadruplet q at time t , T^m_{vqt} is a midline dummy capturing whether the village was treated and T^e_{vqt} is the same variable at endline. [5] η_t and ω_q are time and quadruplet fixed-effects. Standard errors are clustered at the village-level.

Some outcomes were only collected at midline and endline and, for consistency, we estimate the regressions on the midline and endline samples. For outcomes with baseline measures, results are robust to estimating the regressions on the full sample and to estimating ANCOVA regressions. In addition, we obtain similar results when computing p-values through randomization inference (Young, 2018).

4.1 Effects on Political Engagement and Polarization

We begin by examining whether the intervention altered how respondents experience political information. Beyond changing beliefs about specific claims or improving discernment at the individual level, a central goal of the workshops was to reduce perceptions of polarization while fostering productive political engagement.

Measures of affective polarization, including the feeling thermometer, have been developed and validated primarily in the United States (Iyengar et al., 2012; Druckman and Levendusky, 2019) and other established two- or few-party democracies (Gidron et al., 2020; Reiljan et al., 2024), and are difficult to adapt to systems with weaker or less stable party identification (Falkenberg et al., 2026). We instead asked respondents about their perceptions of political division nationally.

Table 1 shows that respondents in our treatment villages were significantly less likely to perceive the Philippines as "more divided than ever" and reported greater interest in politics at both midline and endline. Respondents appear less alienated from politics, not less attentive to it—a pattern rarely observed in the literature. These effects are substantially meaningful and persist over time. It is especially notable because interventions that target polarization can come at the cost of downstream democratic attitudes or political engagement (Voelkel et al., 2023; Fishkin et al., 2021).

The pattern points to a change in how political disagreement is experienced, not a retreat from political life. Changing how people engage with disagreement—not whether they engage at all—is central to how relational approaches can reduce polarization without dampening democratic participation.

[5] We estimate the treatment effects averaged over the three treatment arms. Given that there is no measurable evidence that the two add-ons affected the effectiveness of the base treatment, we pool them in the results reported below.

Table 1: The treatment decreases perceptions of polarization and increases interest in politics

	Country Divided	Interested in Politics
T×Midline	-0.20***	0.075*
	(0.05)	(0.04)
T×Endline	-0.18***	0.13***
	(0.06)	(0.04)
C Mean (Mid)	3.02	1.43
C Mean (End)	3.06	1.41
Obs	3140	3140
R2	0.068	0.048

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable captures, on a 1-10 scale, the degree to which the respondent agrees with the statement *"The country is more divided than ever before"* (Column 1). In Column 2, the dependent variable captures how interested the respondent is about politics (on a 1-4 scale). Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

4.2 Treated individuals change the way they consume news

Next, we examine how the intervention reshaped respondents' information environments. A core premise of the workshops was that polarization is fueled not only by what information people encounter, but by how and where they encounter it. In particular, the intervention worked through community discussion—emphasizing communities as a space for participants to collectively discuss information sources and sharing practices that shape their local political environment.

Consistent with this premise, treated respondents substantially reduced their reliance on major social media platforms for news consumption (Table 2). At endline, treated respondents were 14 percentage-points less likely to use YouTube and 8 percentage-point less likely to use TikTok to get news. Those are large effects given that 77 percent and 35 percent of control group respondents use Youtube and TikTok to get news. These changes are concentrated at endline rather than midline, suggesting durable behavioral adjustment rather than short-term compliance or experimenter demand. These behavioral shifts go beyond platform-level restrictions alone, which have been shown to reduce exposure without producing downstream effects on beliefs or polarization (Ventura et al., 2023).

Importantly, the intervention did not durably change usage of traditional news sources, radio, or television, indicating that respondents did not simply disengage from all media sources.

Table 2: Treated individuals are less likely to use social media to get news

	Do you use the following sources to get news?					
	Facebook	YouTube	TikTok	News Sites	TV	Radio
T×Midline	-0.016	-0.027	-0.024	-0.022	-0.088***	-0.027
	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)
T×Endline	-0.040**	-0.14***	-0.079**	-0.023	-0.018	-0.020
	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
C Mean (Midline)	0.83	0.55	0.24	0.27	0.80	0.27
C Mean (Endline)	0.91	0.77	0.35	0.22	0.76	0.23
Obs	3140	3140	3140	3140	3140	3140
R2	0.033	0.051	0.035	0.033	0.033	0.033

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is a dummy equal to one if the respondent used facebook (Column 1), YouTube (Column 2), TikTok (Column 3), news sites online (Column 4), TV (Column 5) and radio (Column 6) to get news. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

The intervention also altered who respondents talk to about political news (Table 3). Treated respondents became significantly more likely to discuss news within churches and civic groups and less likely to rely on village officials as information intermediaries. This suggests a shift away from official or elite-centered discussion patterns and towards more community-level discussion.

These results are also consistent with a broader pattern of changing local information environments by reducing dependence on social media while strengthening interpersonal discussion within the community.

Table 3: The treatment affects who people talk to about news

	Do you talk about news with?				
	Close Family	Friends in village	Village officials	Members of Church/ social groups	Family outside village
T×Midline	-0.026	-0.027	-0.063**	0.13***	0.076*
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
T×Endline	0.0072	-0.015	-0.054*	0.16***	0.024
	(0.01)	(0.02)	(0.03)	(0.04)	(0.03)
C Mean (Mid)	0.88	0.81	0.71	0.19	0.54
C Mean (End)	0.94	0.88	0.77	0.21	0.61
Obs	3140	3140	3140	3140	3140
R2	0.058	0.038	0.042	0.057	0.041

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is a dummy equal to one if the respondent talk about news with family members living in their village (Column 1), friends in their village (Column 2), village officials (Column 3), member of their church and other social groups (Column 4) friends and relatives living outside their village (Column 5). Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

4.3 Effects extend to the individual's social network

One implication of a relational approach to misinformation is that its effects should extend beyond direct participants. If the intervention operates by reshaping local discussion norms and information practices rather than simply changing individual behavior, then we should observe changes in the broader network and not only the workshop participants.

We find strong evidence that this is the case. Friends and relatives of treated respondents—who did not attend the workshops themselves—also reduced their reliance on YouTube and TikTok for news and were more likely to discuss news in church and civic groups (Table 4). These spillovers are substantial in magnitude, and are in line with the behavioral changes observed among direct participants.

The presence of spillovers also strengthens the interpretation of the media use and discussion results by showing that the intervention altered community-level practices and norms rather than only individual behavior. It also helps rule out explanations based on experimenter demand or short-lived individual learning for these outcomes, as it is unlikely that demand effects would apply to people who never attended a workshop.

Table 4: Spillover effects

	Do you use to get news?			Do you talk about news with?	
	Facebook	YouTube	TikTok	Village officials	Members of Church/ social groups
T×Endline	-0.022	-0.13***	-0.10***	-0.0066	0.046**
	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)
C Mean (Endline)	0.84	0.65	0.32	0.48	0.077
Obs	1200	1200	1200	1200	1200
R2	0.053	0.066	0.046	0.053	0.055

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is a dummy equal to one if the respondent used facebook (Column 1), YouTube (Column 2), TikTok (Column 3) to get news. . The dependent variable is a dummy equal to one if the respondent talk about news with village officials (Column 4), member of their church and other social groups (Column 5). Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

5 Effects on Beliefs and Emotions

Finally, we examine changes in beliefs and emotional responses as downstream consequences of more community-based information environments. Given that the workshops were not designed to correct specific factual claims[6] but to encourage more socially grounded engagement with information, we measure the effect on beliefs in a number of different ways.

5.1 Beliefs

First, treated respondents were less likely to endorse several widely circulating meta-narratives (Table 5). These are long-standing, familiar claims that have circulated and been discussed for some time, and reduced belief in them is consistent with the workshops building resistance to this type of narrative.

We see a similar pattern, more clearly aligned with improved factual reasoning, for beliefs about climate change — an issue where there is broad scientific consensus but significant misinformation online (Table 6). Treated respondents were more likely to believe that temperatures will rise and that typhoons are becoming stronger, with effects persisting at endline. One advantage of using factual claims as a measure is that these beliefs are clear and can be evaluated against established scientific knowledge,

[6]By contrast, some interventions were intended to specifically address issues such as Covid-19 misinformation (Cho et al., 2021; Pennycook et al., 2020; Pennycook and Rand, 2022) or elections-related misinformation (Badrinathan, 2021)

rather than being reliant on partisan messaging.

Table 5: The treatment reduces beliefs in metanarratives

	Sara Duterte	Divorce	China	BGC	K-12
T×Endline	-0.43**	-1.14***	0.24	-0.50***	0.23
	(0.17)	(0.18)	(0.17)	(0.16)	(0.15)
C Mean (End)	4.30	6.58	2.95	4.81	5.87
Obs	1564	1564	1564	1555	1563
R2	0.035	0.097	0.040	0.044	0.043

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is how true they think the following statements are (on a 1-10 scale): *"Sara Duterte should have run for president instead of PBBM's vice president in the 2022 presidential election."* (Column 1), *"The legalization of divorce is bad for Philippine society."* (Column 2), *"Allowing China to operate in the West Philippine Sea means peace for the Philippines."* (Column 3), *"BGC (Bonifacio Global City) does not represent the real Philippines."* (Column 4), *"The implementation of K-12 has ruined the Philippine education system."* (Column 5).. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

Table 6: The treatment affects beliefs about climate change

	Climate is Not Changing	Climate Change is Natural	Temperatures will be hotter	Typhoons Stronger
T×Midline	-0.44**	0.27	0.63***	0.74***
	(0.18)	(0.17)	(0.12)	(0.17)
T×Endline	0.24	-0.053	0.24**	0.32***
	(0.17)	(0.13)	(0.11)	(0.12)
C Mean (Mid)	4.99	5.13	8.02	6.96
C Mean (End)	4.27	4.98	6.88	7.21
Obs	3140	3140	3140	3140
R2	0.044	0.045	0.16	0.055

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is how much they agree with the following statements (on a 1-10 scale): *"The climate is not changing very much."* (Column 1), *"Climate change is a natural process and not caused by humans."* (Column 2), *"Temperatures will be much hotter in the near future."* (Column 3), *"Typhoons are becoming stronger."* (Column 4). Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

At the same time, improved reasoning about metanarratives and scientific facts comes with a trade-off. Because treated respondents rely less on the online platforms where new claims first circulate, they are also less familiar with the specific misinformation in circulation at any given time. Table 7 shows that treated respondents are no better, and if anything worse, at evaluating two stories in circulation at endline. These results suggest that the workshops improved factual reasoning as well as discernment of

familiar or ideologically-coded narratives, even if their reduced exposure to currently circulating misinformation limited their ability to evaluate it.

Taken together, the results point to a form of resilience that is selective rather than general: stronger against familiar, ideologically coded narratives and established facts, weaker against new forms of misinformation that are circulating on the same social media platforms that they are using less.

Table 7: The treatment does not improve respondent's ability to evaluation misinformation circulating online

	EPhillD Story (True)			Chocolate Banned Story (False)		
	Credible	Fact-Check	Share	Credible	Fact-Check	Share
T×Endline	-0.60***	-0.38**	-0.37**	0.36***	-1.07***	0.24*
	(0.12)	(0.17)	(0.18)	(0.12)	(0.21)	(0.13)
C Mean (End)	7.37	5.08	4.87	3.25	4.92	2.79
Obs	1564	1564	1564	1564	1564	1564
R2	0.053	0.046	0.055	0.040	0.073	0.046

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is how much respondents believe that the story on EphillD was credible (Column 1), whether they would fact-check it (Column 2) and whether they would share it (Column 3). In Columns 4-6, the dependent variables are the same but with the story on chocolate being banned in Philippines airport. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

5.2 Emotions

Finally, we examine emotional responses. Consistent with prior work on social media exposure (Allcott et al., 2025), treated respondents report lower levels of anger, annoyance and happiness at endline, alongside a modest increase in nervousness (Table 8). Overall, the emotional profile becomes less reactive and less intense, mirroring patterns observed when individuals reduce exposure to social media platforms. These emotional changes provide further evidence that the intervention altered respondents' day-to-day political information environments.

Table 8: Emotions

	Anger	Fear	Happy	Annoyed	Nervous	Optimistic
T×Midline	-0.12	-0.048	-0.12	-0.041	-0.023	0.089
	(0.13)	(0.14)	(0.10)	(0.13)	(0.12)	(0.11)
T×Endline	-0.46***	0.23	-0.21**	-0.38***	0.34**	-0.038
	(0.16)	(0.15)	(0.10)	(0.14)	(0.13)	(0.13)
C Mean (Mid)	4.96	4.22	8.80	5.14	4.30	8.61
C Mean (End)	4.94	4.12	8.62	5.00	4.04	8.53
Obs	3140	3140	3140	3140	3140	3140
R2	0.036	0.036	0.050	0.045	0.033	0.042

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variables capture the degree to which respondents felt each emotion the previous six months. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

5.3 Results are Consistent Across Baseline Political Beliefs and News Consuption

In terms of heterogeneity, one concern with misinformation and polarization interventions is the potential for differential effects by political beliefs (Jenke, 2024; Bolsen and Druckman, 2018). We do not see evidence that our main effects vary by baseline political beliefs or patterns in their baseline news sources.

These results suggest that reductions in perceived polarization and increases in political interest operate through sustained changes in information environments and social discussion norms, without depending on prior political beliefs or partisanship.

6 Downstream effects on vote choice

We also wanted to see whether the behavioral and attitudinal effects documented above extend to actual electoral outcomes. The intervention substantially reduced reliance on social media for political news and reshaped local discussion networks. These upstream changes suggest a plausible channel through which the effectiveness of digital campaigning and online disinformation could be weakened.

To test this implication, we link village-level treatment assignment to official precinct-level vote returns from the May 2025 Elections (about three to four months after the endline survey and about a year after the workshops). Rather than focusing on a single candidate, we take a broad approach and examine two classes of candidates whose electoral performance is likely to depend on social media exposure: candidates who

invested heavily in Facebook advertising and candidates who were targeted by false or misleading online rumors during the campaign period.

6.1 Candidates who relied heavily on social media received fewer votes

We first analyze electoral results for the five Senatorial candidates with the highest levels of Facebook advertising expenditures.[7] These candidates were disproportionately reliant on social media platforms to reach voters, making their campaigns especially sensitive to changes in voters' media diets.

Table 9 shows that such candidates received significantly lower vote shares in treatment villages. Across specifications, treatment is associated with a reduction of approximately 0.07 to 0.09 percentage points per candidate–precinct pair. Aggregating vote shares for those five candidates at the village-level, the treatment lead to a reduction in vote share of 0.46 percentage-points, corresponding to roughly one-third of a standard deviation. While these effects are modest at the individual precinct level, they are meaningful when aggregated across villages and districts.

Importantly, these results are consistent with the earlier findings on media consumption. Villages exposed to the intervention were less likely to rely on Facebook, YouTube, and TikTok for political news, reducing the reach and effectiveness of digital advertising strategies. The electoral results therefore align with the interpretation that altered information environments weakened candidates' ability to mobilize support through social media.

[7] The data, shared by journalists from the Philippine Center for Investigative Journalism (PCIJ), covers all political ads on Facebook from January 1, 2025 to May 12, 2025 (the day before the elections).

Table 9: Senatorial candidates who spent the most on facebook ads received fewer votes in treatment villages

	Vote Share in the May 2025 Senatorial Elections		
T	-0.071***	-0.092***	-0.46***
	(0.03)	(0.03)	(0.18)
Unit of Observations:	Candidate*Precinct	Candidate*Village	Village
Control Mean	4.27	4.33	21.7
Control Std Dev	1.50	1.60	1.49
Obs	2390	800	160
R2	0.83	0.84	0.61

Notes: Column 1 reports candidate*precinct-level regressions. Column 2 report candidate*village-level regressions. Column report village-level regressions. All regressions include a full set of quadruplet fixed-effects and candidate fixed effects (in Columns 1 and 2). The dependent variable is the candidate vote share in the May 2025 elections. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

6.2 Candidates targeted by negative online rumors in the lead up to the elections received more votes in treatment villages

We next examine candidates who were the subject of negative online rumors during the campaign. Such rumor-based attacks are common in Philippine elections and often rely on rapid dissemination through social media platforms in the days leading up to Election Day.

We leverage the use of such a rumor in our study area. Two days before the elections, rumors started circulating on social media that Ruth and Ranil Hernandez had been disqualified from running for governor and congressman of Laguna by COMELEC (Files, 2025; Rappler, 2025). Both a short video and a post were posted on Facebook pages ("Laguna Facts" and "Laguna Unlocked") emphasizing that a vote for them was a wasted vote. The pages, designed to look like media outlets, had a large number of followers (about 76,000 followers for "Laguna Facts") and the videos and posts were widely shared. A day later COMELEC clarified that the disqualification order wasn't real.[8] Those two candidates ran in our sample municipalities: Ruth Hernandez in all municipalities of the province of Laguna and Ranil in the municipality of Bay (which belongs to Laguna's 2nd district).[9]

[8]Ultimately, Ranil was elected congressman (with a win margin of 2.3 percentage-points) and Ruth was runner-up in the gubernatorial election (with a loss margin of 5.4 percentage-points).

[9]The district consists of the city of Cabuyao and adjacent municipalities of Bay and Los Banos.

As shown in Table 10, candidates targeted by online rumors received approximately 2.8 percentage points more votes in treated villages relative to control villages. While we are unable to test the direct effects of the rumor, the observed pattern suggests that the rumor was partly effective but that our intervention reduced voters' responsiveness to last-minute disinformation campaigns. Indeed, if the rumor was not partly effective in the first place, we would not expect to observe any differences between our treatment and control villages.

Notably, this result does not imply that treated voters became uniformly skeptical of all negative information. Rather, it is consistent with the broader pattern observed throughout the paper: reduced dependence on social media feeds and increased discussion through trusted social networks appear to have dampened the impact of emotionally charged or weakly substantiated claims.

Table 10: Candidates targeted by false online rumors received more votes in our treatment villages

	Vote Share in the May 2025 Elections (Laguna)	
T	2.79**	2.77**
	(1.26)	(1.30)
Observation-level	Candidate*Precinct	Candidate*Village
Control Mean	32.8	32.0
Control Std Dev	13.0	13.2
Obs	353	112
R2	0.85	0.87

Notes: Column 1 reports candidate*precinct-level regressions. Column 2 report candidate*village-level regressions. All regressions include a full set of quadruplet fixed-effects and candidate fixed effects. The dependent variable is the candidate vote share in the May 2025 elections. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

One potential concern with this result is that treatment could have raised support for the Hernandez candidates for reasons entirely unrelated to the rumor. For example, if treatment changed how voters evaluate candidates in general, or shifted support toward incumbents specifically. To address the first possibility, we test whether candidates from the same party as the Hernandezes (LAKAS), who share similar policies, governing styles, and campaign strategies, were also affected by treatment in the same election.[10] To address the second possibility, we test effects on incumbent candidates separately, since both of the Hernandez candidates were incumbent office-holders.

[10]To keep the groups comparable, we focus on candidates for congress, governor, provincial governor, mayor, and vice-mayor.

Table A.6 shows no effect of treatment on vote share for LAKAS candidates or for incumbents. The incumbent estimates are precise enough to exclude an effect of the size we find for the Hernandez candidates (2.8 percentage points); the LAKAS estimates are noisier, and we read them as failing to reject zero rather than as tightly excluding an effect of similar magnitude. These tests are consistent with our result being specific to the rumor rather than reflecting a general change in how treated voters evaluate candidates.

These results complement the earlier evidence on media use and discussion. By reducing exposure to social media and algorithmically targeted content in favor of in-person community-based discussion, the intervention is associated with a weakened effectiveness of social media campaign strategies.

While we do not claim that the intervention deterministically changed vote choices, the election results in our treatment and control areas suggest that changes in information environments can have meaningful political consequences.

7 Mechanisms and Alternative Explanations

The results presented above collectively suggest that the intervention altered how political information circulates within communities, rather than primarily changing individual beliefs about specific claims. Across outcomes—media use, political discussion, perceptions of polarization, and electoral behavior—the effects are consistent with a shift in the local information environment and the social practices that govern information sharing.

One potential concern is that the behavioral and attitudinal effects are entirely due to withdrawing from social media platforms, rather than the content of the workshops themselves. Here, the multiple survey waves is helpful for distinguishing between potential explanations. Political discussion patterns and perceived division are already different in treatment villages at midline, one month after the workshops, while the shift away from social media for political news only becomes significant at endline, nine months later (Tables 1, 3, and 2). This ordering runs against an account in which reduced social media use determines the subsequent changes: discussion partners and perceptions of division change first, and the shift in media consumption for news follows. While our design is not able to establish that the earlier change caused the later one, the order of results is more consistent with an account centered on discussion and community norms than with one centered on social media access alone.

Additionally, while we cannot rule out all alternative explanations, there are several that are inconsistent with the pattern of results. First, experimenter demand or social desirability bias would predict changes in self-reported attitudes without corresponding behavioral or electoral effects. Instead, we observe downstream impacts on media use, social discussion, and vote choice that persist months after the intervention, spillover to untreated individuals (for whom experimenter demand effects are unlikely), and effects reflected in official election returns.

Second, the results are unlikely to be driven by generalized political disengagement or cynicism. Treated respondents report *greater* political interest alongside reduced reliance on social media for political news, a pattern inconsistent with withdrawal from politics. The two findings together suggest that even if social media use for political information is decreasing, individuals are engaging in other forms of political engagement. This is also consistent with respondents' reported openness to new information (Table A.5): agreement that people should always consider new information rises at both midline and endline, which does not fit a story of respondents becoming generally more closed off or cynical. Two other items move toward less openness at endline—greater discomfort with believing something simply because it feels right, and more reluctance to abandon an opinion. However, these two items are more ambiguous in that they can also reflect respondents becoming more selective about the sources they believe. Given that we cannot directly decompose the effects of selectivity from a reduction in openness, we use these results only to show that there is no evidence of attitudes systematically hardening as a result of our treatment, and do not consider them evidence of a fundamental change in general epistemic style.

Third, separate from the issue of whether the treatment worked at all is whether it operated through relational processes or largely through individual-level changes. The presence of spillovers to untreated individuals within treatment villages is hard to explain through individual learning or compliance alone, and instead points to changes in shared information practices. This is consistent with a further pattern: at endline, treated respondents were more likely to say that a person posting misinformation might have been fooled, might be a bot, or was seeking attention, while the item asking whether the poster was trying to mislead others for their own advantage moved at midline but had faded by endline (Table A.3). This shift in how sharers of misinformation are interpreted is more naturally read as a change in social interpretation than as an improvement in individual fact-checking skill. Again, we read these results narrowly and consider them suggestive of a change in perceptions rather than conclusive.

Fourth, the intervention did not target specific candidates, platforms, or campaign mes-

sages, making it unlikely that the electoral effects reflect persuasion against particular actors rather than a reduced responsiveness to social media campaigns or online content more broadly. Candidates who relied more heavily on social media campaigning performed worse in treatment villages, and, in one case of online rumors that we were able to observe, voters in treatment villages were less responsive to a negative rumor against local candidates than voters in control villages. These patterns suggest more fundamental shifts in the political environment in these villages than the workshops would support if the effect operated only through the individual participants.

Fifth, one might worry that the intervention simply made respondents less active on social media in general, rather than specifically changing how they engage with political information. This does not fit the pattern in the data: treated respondents' use of Facebook and TikTok for entertainment did not change significantly at endline, and YouTube fell for entertainment by only about half the magnitude of its decline for news (Table A.2). The changes we document are concentrated in political information, not media use overall.

Last, one further result illustrates the difference between short-term and durable effects, and shows why a single post-treatment survey would have been misleading. Treated respondents' reactions to posts they disagreed with moved in opposite directions across survey waves (Table A.4). At midline, they were more likely to comment, dislike, or report a disagreeable post, and less likely to ignore it; at endline, they were less likely to report, block, or search for more information, and still less likely to ignore it. Had we only observed midline, we might have concluded that the workshops increased confrontational engagement with misinformation online. The endline reversal instead suggests that respondents became more active on these platforms immediately after the workshops, then, seeing no change in outcomes from that activity, moved away from platform-based responses altogether—consistent with the broader shift toward in-person discussion documented above. We cannot fully distinguish this reading from one in which respondents simply disengaged from posts they disagreed with, since we do not observe offline reactions directly, but either way the midline-to-endline reversal underscores the value of tracking outcomes well past the immediate post-treatment period.

Taken together, the evidence points to an intervention that reshaped the relational infrastructure for political information in these communities. While the design does not allow for clean separation of specific causal mechanisms, the results are most consistent with a framework in which misinformation resilience emerges from changes in community-level information practices rather than from individual level correction alone.

8 Conclusion

Social relationships shape both our access to political information and how we interpret it. This paper evaluates a community-based approach to countering misinformation in 160 villages in the Philippines, a context where social media plays a central role in political life and institutional trust is limited. Rather than attempting to improve individual accuracy through correction or fact-checking, the intervention was designed to reshape local information environments by encouraging collective reflection, discussion, and responsibility in information sharing.

Treated respondents discussed the news more with church and civic groups and less with village officials, relied less on social media, and perceived the country as less divided. Importantly, these effects did not come at the cost of disengagement from politics: respondents reported more interest in politics, not less.

These changes even reached people who never attended the workshops: named discussion partners of treated respondents showed the same shifts in media use and discussion.

These effects also translated to election results almost a year later. In the May 2025 senatorial election, candidates who relied heavily on Facebook advertising underperformed in treatment villages, and voters in treated villages were less susceptible to online rumors targeting local candidates.

Measuring discernment across different domains illustrates the trade-offs associated with the relational approach. The workshops helped treated individuals assess familiar, ideologically-coded narratives and well-established facts. By contrast, the workshops did not help their ability to evaluate current misinformation actively circulating at the time of the survey. This is consistent with reduced exposure: treated respondents relied less on the platforms where such claims tend to circulate.

Taken together, the results suggest that misinformation is not only an informational problem but a social one. When political information is embedded in relationships of trust and belonging, interventions aimed only at individual accuracy leave the social incentives for information sharing unchanged. Targeting a community's information environment—its norms and practices around sharing—can build resilience to misinformation.

At the same time, communities are not resilient or protective by default. The same social networks that facilitate collective resilience to misinformation can just as plausibly spread and sustain misinformation. Whether community discussion filters misin-

formation or transmits it depends in large part on the community's incentives for information sharing.

Another important condition is the quality of the alternative information sources. This is a scope condition on our results as much as a target for future research. Community discussion only outperforms a social media platform as a source of political information when the platform itself is a source of misinformation. In a setting with weak content moderation and heavy reliance on algorithmically ranked feeds, as in the Philippines, the marginal neighbor may simply be a better source than the marginal post. The reverse can also hold: where traditional media and local institutions are state-controlled or otherwise compromised, social media may be a comparatively freer channel, and less reliance on it would not be an improvement. Which alternative is better—the online or the in-person social network—is an empirical question that depends on the setting, not a general property of either one.

In terms of policy implications, the results suggest that community-based workshops can complement efforts on platform regulation and content moderation, particularly in settings where social media dominates the flow of political information. An additional advantage of relational interventions is the ability to scale through the social networks they target: a fixed, modest investment per village produced effects that reached well beyond the people who attended and were still evident in election outcomes almost a year later.

References

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang, "Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic," *Journal of Public Economics*, 2020, *191*, 104254.

_ , Matthew Gentzkow, Benjamin Wittenbrink, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Sandra Gonzalez-Bailon, Andrew M Guess, Young Mie Kim et al., "The Effect of Deactivating Facebook and Instagram on Users' Emotional State," Working Paper 33697, National Bureau of Economic Research 2025.

Amar, Priyadarshi, Sumitra Badrinathan, Simon Chauchard, and Florian Sichart, "Countering Misinformation Early: Evidence from a Classroom-Based Field Experiment in India," *American Political Science Review*, 2025, p. 1–21.

Arugay, Aries A. and Justin Keith A. Baquisal, "Mobilized and Polarized: Social Media and Disinformation Narratives in the 2022 Philippine Elections," *Pacific Affairs*, 2022, *95* (3), 549–573.

_ and Maria Elize H. Mendoza, "Digital Autocratization and Disinformation in Philippine Elections," 2024.

Badrinathan, Sumitra, "Educative interventions to combat misinformation: Evidence from a field experiment in India," *American Political Science Review*, 2021, *115* (4), 1325–1341.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson, "The Diffusion of Microfinance," *Science*, 2013, *341* (6144), 1236498.

Banerjee, Abhijit V., Arun Chandrasekhar, Esther Duflo, and Matthew O. Jackson, "Using gossips to spread information: theory and evidence from two randomized controlled trials," *NBER WP 20422*, 2018.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler, "The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections," *Journal of Experimental Political Science*, 2023, *10* (1), 34–49.

Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield, "Interventions to counter misinformation: Lessons from the Global North and applications to the Global South," *Current Opinion in Psychology*, 2024, *55*, 101732.

Bolsen, Toby and James N. Druckman, "Do partisanship and politicization undermine the impact of a scientific consensus message about climate change?," *Group Processes & Intergroup Relations*, 2018, *21* (3), 389–402.

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, Shelley Liu, and John Marshall, "Sustaining Exposure to Fact-Checks: Misinformation Discernment, Media Consumption, and Its Political Implications," *American Political Science Review*, 2025, *119* (4), 1864–1887.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, July 2017, *114* (28), 7313–7318. Publisher: Proceedings of the National Academy of Sciences.

Cho, Yoonyoung, Cesi Cruz, Julien Labonne, Kevin Lim, Sergio Piza, and Ben Rivera, "Reducing Vaccine Hesitancy in the Philippines: Findings from a Survey Experiment," Technical Report, World Bank 2021.

Chopra, Felix, Ingar Haaland, and Christopher Roth, "The Demand for News: Accuracy Concerns Versus Belief Confirmation Motives," *The Economic Journal*, 2024, *134* (661), 1806–1834.

Chua, Yvonne T., "Filipino Concern over Online Disinformation Hits Record High—Digital News Report 2025," VERA Files, June 17, 2025 2025.

_ , "Philippines," in Nic Newman, Amy Ross Arguedas, Craig T. Robertson, Rasmus Kleis Nielsen, and Richard Fletcher, eds., *Reuters Institute Digital News Report 2025*, Oxford: Reuters Institute for the Study of Journalism, 2025, pp. 148–149.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, 2021, *118* (9), e2023301118.

Druckman, James N. and Matthew S. Levendusky, "What Do We Measure When We Measure Affective Polarization?," *Public Opinion Quarterly*, 2019, *83* (1), 114–122.

Dugas, Michelle, Daniel Pinzón, JungKyu Rhys Lim, Renos Vakis, Zeina Afif, Takahiro Hasumi, and Diya Elfadel, "Find the Fake: Boosting Resistance to Health Misinformation in Jordan with a WhatsApp Chatbot Game," Policy Research Working Paper 10933, World Bank, Washington, DC September 2024.

Falkenberg, Max, Matteo Cinelli, Alessandro Galeazzi, Christopher A. Bail, Rosa M. Benito, Axel Bruns, Anatoliy Gruzd, David Lazer, Jae Kook Lee, Jennifer McCoy, Kikuko Nagayoshi, David G. Rand, Antonio Scala, Alexandra A. Siegel, Sander

van der Linden, Onur Varol, Ingmar Weber, Magdalena Wojcieszak, Fabiana Zollo, Andrea Baronchelli, and Walter Quattrociocchi, "Toward an Integrated and Globally Equitable Understanding of Political Polarization," *Nature Communications*, 2026, *17*, 8338.

Files, Vera, "FACT CHECK: Hernandez couple of Laguna NOT disqualified from local elex," *https://verafiles.org/articles/fact-check-hernandez-couple-of-laguna-not-disqualified-from-local-elex*, 2025.

Fishkin, James, Alice Siu, Larry Diamond, and Norman Bradburn, "Is deliberation an antidote to extreme partisan polarization? Reflections on "America in One Room"," *American Political Science Review*, 2021, *115* (4), 1464–1481.

Flynn, D. J., Brendan Nyhan, and Jason Reifler, "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics," *Political Psychology*, 2017, *38* (S1), 127–150.

Gentzkow, Matthew and Jesse M. Shapiro, "Ideological Segregation Online and Offline *," *The Quarterly Journal of Economics*, 11 2011, *126* (4), 1799–1839.

Gidron, Noam, James Adams, and Will Horne, *American Affective Polarization in Comparative Perspective* Elements in American Politics, Cambridge: Cambridge University Press, 2020.

Gottlieb, Jessica, Claire L. Adida, and Richard Moussa, "Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire," December 2022.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India," *Proceedings of the National Academy of Sciences*, 2020, *117* (27), 15536–15545.

Huckfeldt, R. Robert and John Sprague, *Citizens, Politics, and Social Communication: Information and Influence in an Election Campaign*, Cambridge: Cambridge University Press, 1995.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes, "Affect, Not Ideology: A Social Identity Perspective on Polarization," *Public Opinion Quarterly*, 2012, *76* (3), 405–431.

Jenke, Libby, "Affective Polarization and Misinformation Belief," *Political Behavior*, June 2024, *46* (2), 825–884.

Katz, Elihu and Paul F. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications*, Glencoe, IL: Free Press, 1955.

Kobayashi, Tetsuro, Zhifan Zhang, and Ling Liu, "Is Partisan Selective Exposure an American Peculiarity? A Comparative Study of News Browsing Behaviors in the United States, Japan, and Hong Kong," *Communication Research*, Oct 2024.

Kunda, Ziva, "The case for motivated reasoning," *Psychological Bulletin*, 1990, *108* (3), 480–490.

Larreguy, Horacio and Pia J. Raffler, "Accountability in Developing Democracies: The Impact of the Internet, Social Media, and Polarization," *Annual Review of Political Science*, 2025, *28*, 413–434.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al., "The science of fake news," *Science*, 2018, *359* (6380), 1094–1096.

Levy, Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, March 2021, *111* (3), 831–870.

Li, Jianing and Michael W. Wagner, "The Value of Not Knowing: Partisan Cue-Taking and Belief Updating of the Uninformed, the Ambiguous, and the Misinformed," *Journal of Communication*, 2020, *70* (5), 646–669.

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

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand, "Shared partisanship dramatically increases social tie formation in a Twitter field experiment," *Proceedings of the National Academy of Sciences*, 2021, *118* (7), e2022761118.

Nyhan, Brendan and Jason Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 2010, *32* (2), 303–330.

Ong, Jonathan Corpus, "Philippine Elections 2022: The Dictator's Son and the Discourse around Disinformation," *Contemporary Southeast Asia*, 2022, *44* (3), 396–403.

_ **and Ross Tapsell**, "Demystifying disinformation shadow economies: fake news work models in Indonesia and the Philippines," *Asian Journal of Communication*, 2022, *32* (3), 251–267.

Parihar, Sushmeena, Trisha Harjani, Priyank Mathur, Beth Goldberg, S van der Linden, and J Roozenbeek, "Countering misinformation in India through prebunking," *PsyArXiv Preprints*, 2025.

Pennycook, Gordon and David G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, 2019, *188*, 39–50.

_ **and** _ , "Accuracy Prompts Are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation," *Nature Communications*, 2022, *13* (1).

_ **, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand**, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science*, 2020, *31* (7), 770–780.

Peterson, Erik and Ali Kagalwala, "When unfamiliarity breeds contempt: How partisan selective exposure sustains oppositional media hostility," *American Political Science Review*, 2021, *115* (2), 585–598.

Porter, Ethan and Thomas J. Wood, "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom," *Proceedings of the National Academy of Sciences*, September 2021, *118* (37), e2104235118.

Rappler, "FACT CHECK: Laguna governor, district representative not disqualified by Comelec," *https://www.rappler.com/newsbreak/fact-check/philippines-elections-laguna-governor-district-representative-not-disqualified-comelec/*, 2025.

Reiljan, Andres, Diego Garzia, Frederico Ferreira da Silva, and Alexander H. Trechsel, "Patterns of Affective Polarization toward Parties and Leaders across the Democratic World," *American Political Science Review*, 2024, *118* (2), 654–670.

Taber, Charles S. and Milton Lodge, "Motivated Skepticism in the Evaluation of Political Beliefs," *American Journal of Political Science*, 2006, *50* (3), 755–769.

Tapsell, Ross, "Social Media and Elections in Southeast Asia: The Emergence of Subversive, Underground Campaigning," *Asian Studies Review*, 2021, *45* (1), 117–134.

Ventura, Tiago, Sayan Majumdar, Jonathan Nagler, and Joshua A. Tucker, "WhatsApp increases exposure to false rumors but has limited effects on beliefs and polarization," 2023. SSRN Working Paper. doi: 10.2139/ssrn.4457400.

Voelkel, Jan G., Johanna Chu, Michael N. Stagnaro, Joseph S. Mernyk, Chrystal Redekopp, Sophia L. Pink, James N. Druckman, David G. Rand, and Robb Willer, "Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes," *Nature Human Behaviour*, 2023, *7*, 55–64.

Vosoughi, Soroush, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science*, 2018, *359* (6380), 1146–1151.

Wardle, Claire and Hossein Derakhshan, "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making," Technical Report, Council of Europe 2017.

Yarchi, Moran, Christian Baden, and Neta Kligler-Vilenchik, "Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media," *Political Communication*, 2021, *38* (1–2), 98–139.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *The Quarterly Journal of Economics*, 2018, *134* (2), 557–598.

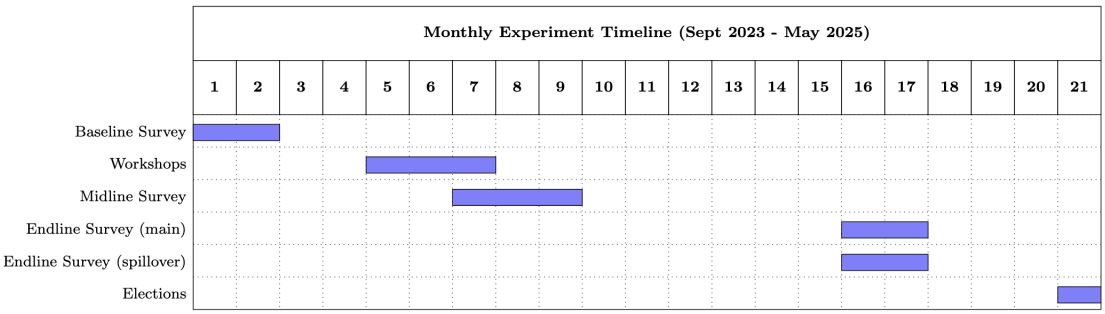


Figure 1: Timeline

From Correction to Connection: Relational Approaches to Countering Misinformation

ONLINE APPENDIX
NOT FOR PUBLICATION

Table A.1: List of Sample Municipalities.

Province	Municipality	Number of Sample Villages
Laguna	Alaminos	12
Laguna	Bay	12
Laguna	Calauan	12
Laguna	Pila	12
Laguna	Santa Cruz	20
Laguna	Victoria	8
Laguna	Nagcarlan	24
Quezon	Candelaria	16
Quezon	Dolores	12
Quezon	Sariaya	32

Table A.2: More muted effects on use of sources for entertainment

	Do you use for entertainment?			Do you talk about entertainment with?	
	Facebook	YouTube	TikTok	Village officials	Members of Church/ social groups
T×Endline	-0.012	-0.068***	-0.039	-0.054*	0.16***
	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
C Mean (Endline)	0.94	0.88	0.45	0.77	0.21
Obs	3140	3140	3140	3140	3140
R2	0.027	0.034	0.022	0.042	0.057

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable is a dummy equal to one if the respondent used facebook (Column 1), YouTube (Column 2), TikTok (Column 3) to get news. . The dependent variable is a dummy equal to one if the respondent talk about news with village officials (Column 4), member of their church and other social groups (Column 5). Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

Table A.3: The treatment changes the way individuals who post misinformation are perceived

	Attention Seeking	Fooled	Bot	Trying Mislead
T×Midline	0.23***	-0.0039	-0.099	0.18***
	(0.05)	(0.05)	(0.07)	(0.04)
T×Endline	0.36***	0.34***	0.26***	0.059
	(0.06)	(0.06)	(0.06)	(0.06)
Control Mean (Midline)	3.10	3.06	2.84	3.24
Control Mean (Endline)	2.95	2.41	2.27	3.20
Obs	2968	2968	2968	2968
R2	0.072	0.095	0.068	0.059

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variables capture responses to the question "*When you see misinformation on social media, which of the things below is likely to be true of the person who posted it?*" In Column 1, the answer is "they're seeking attention". In Column 2, the answer is "they might have been fooled". In Column 3, the answer is "they might be a bot". In Column 4, the answer is "They're trying to mislead others for their own advantage". All answers are on a 1-4 scale. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

Table A.4: The treatment changes the way respondents reac to misinformation posts

	Write Comment	Attack	Dislike	Report	Block	Search Info	Ignore
T×Midline	0.27***	-0.0065	0.20***	0.20***	0.040	0.056	-0.17***
	(0.05)	(0.02)	(0.05)	(0.04)	(0.04)	(0.07)	(0.06)
T×Endline	-0.019	-0.036	0.069	-0.12***	-0.15***	-0.34***	-0.14**
	(0.04)	(0.02)	(0.05)	(0.03)	(0.04)	(0.07)	(0.06)
C Mean (Mid)	1.56	1.19	1.42	1.43	1.52	2.34	3.23
C Mean (End)	1.25	1.13	1.29	1.20	1.35	2.17	3.34
Obs	2966	2966	2966	2966	2966	2966	2967
R2	0.051	0.021	0.049	0.074	0.053	0.064	0.053

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variables capture responses to the question "*When you see a post on social media that you don't agree with, how likely are you to take the following actions?* . In Column 1, the answer is "Write a comment disagreeing with the content". In Column 2, the answer is "Write a comment attacking the individual who posted it". In Column 3, the answer is "Hit dislike immediately". In Column 4, the answer is "Report the post". In Column 5, the answer is "Block the content creator". In Column 6, the answer is "Look for more information about the topic". In Column 7, the answer is " Ignore it/Scroll down". All answers are on a 1-4 scale. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

Table A.5: Openness

	Consider New Information	Comfortable	Too Important	Conflicts	Confusing	Against
T×Midline	0.16***	-0.20***	-0.14***	-0.063	0.25***	0.16**
	(0.04)	(0.05)	(0.04)	(0.06)	(0.05)	(0.06)
T×Endline	0.12***	0.12**	0.11***	0.032	0.055	-0.0043
	(0.04)	(0.05)	(0.04)	(0.06)	(0.04)	(0.03)
C Mean (Mid)	3.27	2.81	3.06	2.29	2.82	3.21
C Mean (End)	3.23	2.66	2.86	2.35	2.97	3.34
Obs	3139	3139	3139	3139	3139	3140
R2	0.069	0.043	0.043	0.064	0.043	0.030

Notes: Results from individual-level regressions. All regressions include a full set of quadruplet fixed-effects. The dependent variable capture the degree to which the respondent agree with a number of statements (on a 1-4) scale. In Column 1, the statement is "A person should always consider new information.". In Column 2, the statement is "If something I think feels right then I am comfortable, whether it is true.". In Column 3, the statement is "Certain opinions are just too important to abandon no matter what else you learn." In Column 4, the statement is "One should disregard information that conflicts with your current opinions.". In Column 5, the statement is "Considering too many different opinions can make you more confused.". In Column 6, the statement is "People should take into consideration information even if it goes against their opinions.". Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.

Table A.6: Incumbents and Candidates from the same party as those targeted by the rumors do NOT receive fewer votes

	Vote Share in the May 2025 Elections			
	Lakas Candidates		Incumbents	
T	0.57	-0.34	-0.39	-1.03
	(1.10)	(1.33)	(0.53)	(0.74)
Observation-level	Candidate* Precinct	Candidate* Village	Candidate* Precinct	Candidate* village
Control Mean	60.1	62.7	53.2	57.1
Control Std Dev	28.3	28.4	37.5	37.3
Obs	1171	404	2002	680
R2	0.93	0.92	0.96	0.96

Notes: Columns 1 and 3 reports candidate*precinct-level regressions. Columns 2 and 4 report candidate*village-level regressions. All regressions include a full set of quadruplet fixed-effects and candidate fixed effects. The dependent variable is the candidate vote share in the May 2025 elections. Standard errors (in parentheses) clustered by village. * p < .10, ** p < 0.05, *** p < .01.