Title
Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire

Authors
Gottlieb, Jessica
Adida, Claire L.
Moussa, Richard

Publication Date
2022-12-22

DOI
10.26085/C3Q30T

Series Name: WPS
WPS Paper No. 217
Issue Date: 12/15/22

Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire

Jessica Gottlieb, Claire L. Adida, and Richard Moussa

CEGA

Center for Effective Global Action

Working Paper Series

Center for Effective Global Action

University of California

eScholarship
University of California

Recommended Citation:

Gottlieb, Jessica; Adida, Claire L.; Moussa, Richard. (2022). Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire. Working Paper Series No. WPS-217. Center for Effective Global Action. University of California, Berkeley, Text. https://doi.org/10.26085/C3Q30T

Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire*

Jessica Gottlieb† Claire L. Adida‡ Richard Moussa§

December 14, 2022

Abstract

Misinformation has deleterious and potentially destabilizing effects on democracy. As a result, scholars and practitioners alike are investigating strategies to reduce the belief in and dissemination of misinformation. A common strategy is a digital literacy intervention to increase individual capacity to identify misinformation online. This, we argue, ignores identity-based motivations to consume biased media. We offer a theoretical framework that highlights the limitations of strategies that ignore individuals' directional motives. We propose three interventions that leverage insights on how social identity shapes behavior, and test each with an information experiment in Côte d'Ivoire, a polarized country. We find that a standard digital literacy intervention fails to curb the belief in and spread of misinformation, while our social-identity based interventions limited both. Our findings confirm that misinformation spreads at least in part because individuals are motivated to consume biased media, and caution against strategies that ignore these directional motives.

*We thank Alex Coppock and Guy Grossman for feedback on our research design; Andrew Little for especially insightful comments on interpreting results; and participants at the 2022 Abidjan EGAP Policy Event, Oxford's Nuffield College Political Science Seminar, and the UT-Austin conference on authoritarian regimes and democratic backsliding. We are also grateful to Linh Le for research assistance and to our partners at the National Democratic Institute for a fruitful collaboration. This study is pre-registered as EGAP Registration 20211117AA at https://osf.io/n28wr. It received IRB approval from UC San Diego and University of Houston.

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

‡Professor, UC San Diego

§Assistant Professor, Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSEA)

1 Introduction

Just over a year ago, violent clashes erupted in Abidjan, the capital of Côte d'Ivoire, killing one and injuring many more individuals of Nigerien origin. The cause? A video claiming to show Ivoirian migrants in neighboring Niger suffering from attacks themselves. In fact, the video was a 2019 clip from Nigeria showing the country's military arresting members of the terrorist group Boko Haram. It was fake news, with fatal implications.[1]

Misinformation, and its spread through social media, exacerbate affective polarization – or the perceived social distance between groups (Iyengar, Sood and Lelkes, 2012; Lelkes, 2016; Suhay, Bello-Pardo and Maurer, 2018; Tucker et al., 2018). As a result, scholars are interrogating the role of social media in destabilizing democratic politics and fomenting conflict (Tucker et al., 2017; Lorenz-Spreen et al., 2022). If political polarization has deleterious effects on democracy – by eroding civil discourse and political compromise thus escalating gridlock and political brinkmanship (McCoy, Rahman and Somer, 2018; Tucker et al., 2018), or by leading citizens to prioritize partisanship over democratic principles thus weakening accountability (Graham and Svolik, 2020) – then social media's impact on polarization could contribute to democratic erosion.[2] In young democracies with recent history of civil conflict, such as Côte d'Ivoire, the effect may be even more destabilizing.

Côte d'Ivoire is a polarized country; its two civil wars pitted Southerners against Northerners in a battle to define the true Ivoirian identity (Ivoirité) (Konate, 2004). At the same time, polarization is exacerbated by a number of other identities that cleave along the same lines: Southerners also tend to be Christian, ethnic Akans, and supporters of the opposition party PDCI; and Northerners also tend to be Muslim, ethnic Mandinké or Burkinabé, and supporters of the ruling party RHDP. Against this backdrop, the country has an internet penetration rate of approximately 36%, with

[1] See `http://apanews.net/en/news/icoast-ethnic-clashes-caused-by-fake-news-leave-10-people-wounded`.

[2] While recent studies call into question the immediate causal relationship between affective polarization and democratic attitudes (Broockman, Kalla and Westwood, 2022), severe polarization may nevertheless undermine the quality of democracy in the longer run.

more than 23% of its population using social media, a number that is growing rapidly every year (by 8.5% between 2021 and 2022).[3]

Scholars and practitioners alike see this combination as a potentially explosive one.[4] A recent study by the National Democratic Institute (NDI) shows that rampant misinformation characterizes Facebook's most widely-shared posts in Côte d'Ivoire, including narratives about assassination attempts against political leaders (Gatewood et al., 2020). Similarly, the Center for Democracy and Development reports that fake news in Côte d'Ivoire, which spreads primarily via WhatsApp and Facebook, amplifies political tensions and social divisions particularly around presidential and legislative elections.[5] This dangerous pattern is not unique to Côte d'Ivoire: in Kenya in September 2021, a Mozilla Foundation investigation uncovered a coordinated misinformation campaign to undermine Kenya's High Court – at a time when the court was reviewing the president's controversial initiative for constitutional reform.[6] Consequently, prominent democracy-promoting organizations like NDI are now focusing much of their own programming on combating online misinformation and promoting digital literacy. At the same time, standard digital literacy programs have a weak track record of success in developing countries (Badrinathan, 2021), motivating alternative approaches to reducing online misinformation.

This paper develops a theoretical framework that helps us understand the conditions under which digital literacy interventions effectively reduce misinformation uptake and dissemination. It focuses on the demand for misinformation: individuals may be motivated to consume biased media because of the psychological benefits it confers. This creates a challenge to standard interventions that attempt to counter misinformation. Without additional interventions that explicitly address these directional motives, we argue, standard approaches to countering misinformation are unlikely

[3] See https://datareportal.com/reports/digital-2022-cote-divoire.

[4] See https://africanarguments.org/2022/01/the-genocide-that-never-was-and-the-rise-of-fake-news-in-cote-divoire/.

[5] See https://www.developmentdiaries.com/2022/01/ivory-coast-how-fake-news-impacts-nation-cdd/.

[6] See https://foundation.mozilla.org/en/blog/fellow-research-inside-the-shadowy-world-of-disinformation-for-hire-in-kenya/.

to work.

We draw from rich literatures across the social sciences to propose three alternative approaches to countering misinformation, all of which leverage the political salience of social identities in a polarized context. First, we draw from scholarship in social psychology and more recently political science on the robust effects of empathy on outgroup inclusion (Kalla and Broockman, 2020). Second, we look to the compelling findings about the power of social norms to reduce prejudice (Paluck and Green, 2009; Paluck, Shepherd and Aronow, 2016). Finally, we draw from the literature on elite endorsements to test whether individuals seeking to enhance their own status may look to cues from popular elites (McClendon, 2018; Bullock, 2020). All three approaches aim to reduce misinformation uptake by reducing affective polarization or the motivation to assimilate biased information. We use these three approaches to develop an empathy intervention, a social norms intervention, and a popular elite intervention to complement a standard digital literacy intervention.

Our empirical strategy was to partner with an international non-governmental organization working on democracy promotion in Côte d'Ivoire to develop interventions that both operationalized our theoretical intuitions and reflected ongoing efforts by actual practitioners on the ground. We embedded treatment and placebo interventions in a survey experiment following a factorial design: respondents received either a digital literacy or a placebo financial literacy intervention (by video); and for each, we either added nothing or one of three social-identity based interventions as described above (by audio or video). We recruited young Ivoirians in the country's urban center, Abidjan, to enroll them in an online program evaluation. And we administered a two-wave panel survey, implementing a baseline questionnaire, the treatment, and a manipulation check in wave 1, and measuring our outcomes two to six weeks later in wave 2.

Our findings are fourfold. First, we indeed find that the standard digital literacy intervention by itself has no effect on misinformation uptake and sharing, corroborating the argument that digital literacy interventions alone cannot change the way individuals consume and share information in a polarized context. Second, we find that all three of our social-identity based interventions have

significant effects on misinformation identification and sharing. Specifically, we find a positive effect of our empathy intervention on misinformation identification, a weak but (unexpected) negative effect of our norm intervention on misinformation identification, and a robust negative effect of our popularity intervention on misinformation sharing. Finally, and counter to expectations, we see no evidence that any of these effects are occurring through a decrease in affective polarization.

Our results contribute to a rich literature on social media and polarization by applying social scientific insights to investigate the effectiveness of interventions to reduce online misinformation. They corroborate our intuition about the ineffectiveness of standard digital literacy interventions, which are naive to the powerful effect of social identities on information uptake. Indeed, this study provides compelling and in-the-wild empirical grounding for social science research highlighting the role that social identities can play in political persuasion. Additionally, it can inform interventions by international non-government organizations working to reduce political polarization and conflict.

Our results also contribute to the ongoing debate about the role that motivated reasoning plays in information uptake. Social scientists disagree about the pervasiveness of motivated reasoning: some have argued that individuals tend not to integrate new information when it conflicts with directional goals (Kahan, 2016). This is especially salient in polarized contexts where individuals derive self-esteem from their social identities (Ehret, 2021). Yet others have questioned the extent to which motivated reasoning shapes information uptake, showing instead that individuals tend to update their beliefs in a manner that is consistent with Bayes' Rule (Coppock, 2016), and that many studies purporting to show results that are consistent with motivated reasoning are actually also consistent with Bayesian updating (Little, 2021). By directly manipulating directional bias, our study conclusively shows that motivated reasoning is at play - and how to alleviate it.

2 Theoretical Framework

Media bias and misinformation persist for both supply-side and demand-side reasons (Gentzkow, Shapiro and Stone, 2015). Here, we focus on the demand-side: the preference among consumers to assimilate biased media because it offers greater psychological utility. This behavior is consistent with motivated reasoning or the impetus, in taking up new information, to maximize the dual goals of accuracy and group identity affirmation (Taber and Lodge, 2006). If assimilating new information offers a psychological benefit because it is consistent with prior ingroup beliefs or makes the consumer feel better as a member of that ingroup (Tajfel and Turner, 1986; Kahan, 2016), then we may observe the uptake of misinformation even when information-consumers know it is false. Or, as Peterson and Iyengar (2021) show, partisans may sincerely believe misinformation when they are motivated to uncritically accept information favorable to their side and ignore factual counterarguments.

We propose that these patterns of motivated behavior pose a special challenge to the standard interventions that attempt to counter misinformation. Misinformation interventions can be categorized into two types that intervene either before or after a consumer encounters the misinformation. Ex post interventions attempt to correct misinformation by providing fact checks, while ex ante interventions attempt to inoculate against or prevent the consumer from believing future encounters with misinformation. Both types of interventions implicitly assume that consumers value accuracy, or at least value it relatively more than some other psychological utility such as the affirmation of their ingroup identity.

2.1 Interventions to Correct Misinformation

Empirically, scholars have shown that purely informational interventions rarely change attitudes (Flynn, Nyhan and Reifler, 2017; Hopkins, Sides and Citrin, 2019; Nyhan et al., 2020), though factual corrections do lead consumers to update beliefs about whether a specific piece of information

is true or false – even if they may not subsequently shift related political attitudes. For instance, a recent meta-analysis found that fact-checking leads consumers to hold more correct beliefs when faced with political misinformation (Walter et al., 2020). This consensus finding largely comes from studies in developed country contexts, however, and Batista et al. (2022) show that fact-checking corrections in Brazil are ineffective at reducing misinformation (though see Bowles et al. (2022) for a successful fact-checking intervention in South Africa).

The social media news environment also poses a special problem. "As fact-checking standards online are lax, low entry barriers together with the unprecedented speed with which users can share content on social media could lead to a spread of misinformation and fake news, ultimately increasing political misperceptions" (Zhuravskaya, Petrova and Enikolopov, 2020). While another recent study found that misinformation corrections can work even in a noisy social media environment (Porter and Wood, 2022), such corrections are nearly impossible to implement when much of the online misinformation is disseminated via Whatsapp's closed-messaging environment where there is no moderator that can insert factual corrections as there might be on other platforms like Twitter or Facebook. Guess and Lyons (2020) discuss how relatively little is known about misinformation on this platform and its relatively greater importance among contexts in the Global South.

Scholars and policymakers have long recognized the limitations of fact-checking or misinformation corrections. Not only is fact-checking not always feasible or effective, but the mere fact of being exposed to misinformation can already take effect before a post hoc correction is put in place. And effects may linger even after misinformation is corrected – what Lewandowsky et al. (2012) call the continued influence effect. Instead, some scholars have long advocated "inoculation" against misinformation, or that it is more advantageous to expose people to a small dose of misinformation (alongside preemptive refutation) than to try to undo the misinformation attack ex post (McGuire, 1964; Traberg, Roozenbeek and van der Linden, 2022).

Interventions to combat misinformation in the developing world rely largely on digital literacy or educational programs that aim to inoculate information-consumers against fake news by making

them more aware of their media environment, warning against the prevalence of misinformation, and providing strategies for identifying fake news. Studies in the US, and one in South Africa, have shown some promising evidence of the efficacy of such interventions (Tully, Vraga and Bode, 2020; Vraga, Bode and Tully, 2020; Cook, Lewandowsky and Ecker, 2017; Bowles et al., 2022). But a recent intervention in India elicited no effect, made all the more surprising by the intervention's intensity – a one-hour in-person media literacy training (Badrinathan, 2021). Another recent study of a media literacy intervention in both the US and India provides some insight into this disparity. While the US intervention worked among a representative sample of participants, the Indian intervention worked only among the highly educated (Guess et al., 2020).

2.2 Motivated Reasoning and Misinformation

Given mixed findings on the effect of interventions designed to improve consumers' *capacity* to reject misinformation, we posit that consumers may additionally require a reduction in their *motivation* to take up misinformation. This should be especially true in a context like Côte d'Ivoire where group identity is particularly salient and socio-political issues polarize around those identities. Here, we draw from rich literatures on persuasion and prejudice-reduction to make predictions about the kinds of interventions that will effectively demotivate consumers to believe biased or false information.

In polarized contexts, social identities become politically and socially salient: individuals identify with certain social categories and are motivated to affirm and reaffirm these social identities. Furthermore, one's ingroup identity is partially defined in opposition to the outgroup so high levels of animus toward the outgroup can inspire increased identification with the ingroup (Mason, 2018).

While the term polarization can also describe ideological distance between two groups, this study is most concerned with affective polarization or the extent to which individuals express antipathy toward members of outgroups, and/or show affinity for members of their own group. Our understanding of the phenomenon is influenced by work on partisan polarization in the US that

conceptualizes affective polarization as the result of individuals internalizing their partisanship as a social identity (Huddy, Mason and Aarøe, 2015; Iyengar et al., 2019). This generates strong ingroup preferences and outgroup bias that can be distinct from ideological preferences (Mason, 2018). In our empirical context, this internalization of partisanship as a social identity makes sense as identity is an explicit driver of political behavior (McCauley, 2017). Where political polarization is driven in large part by social identity attachments, social-identity-based interventions to reduce affective polarization may more effectively move individuals to change the way they process new information.

A first such intervention relies on recent findings about the effects of empathy and perspective-taking. The ability to imagine oneself in another person's shoes can reshape the boundaries between ingroup and outgroup, triggering what social psychologists call self-other merging (Cialdini et al., 1997). By doing so, an intervention that encourages individuals to empathize with others might improve their attitudes about the outgroup and lower polarization (Voelkel et al., 2022), thereby decreasing an individual's motivation to stick to misinformation that affirms their social identity.

A second intervention draws instead on the compelling findings about the power of social norms. Social scientists have long argued that behavior may reflect norms, or the belief that others in a peer group adhere to the same set of preferences, beliefs, attitudes and behaviors. For example, Black Americans often sacrifice personal self-interest for racial group interest because of strong social norms within the Black American community (White, Laird and Allen, 2014); and interpersonal conflict among middle school students in the United States dropped significantly due to an intervention in which socially popular students were randomly assigned to anti-conflict training and to become the public face of opposition to conflict (Paluck, Shepherd and Aronow, 2016). Across a diversity of contexts, social norms powerfully shape how individuals behave. A key reason for this is the expectation of ingroup policing – or sanctioning from fellow group members for deviation from the group norm (Fearon and Laitin, 1996; Habyarimana et al., 2007). Consistent with these results, an intervention that changes people's perceptions about their own group's social

norms about the outgroup – particularly one that highlights a diversity of experiences with and opinions about outgroup members, thus reducing the threat of in-group policing – might reduce prejudice, polarization, and the motivation to take up misinformation.

Finally, the literature on elite endorsements suggests that elites might have a role to play in shaping individual attitudes (Boudreau, 2019; Pink et al., 2021). This may be the case if indeed individuals care not just about their social identities, but their positions within their social groups (McClendon, 2018). Individuals seeking to enhance their own status may look to and respond to cues from popular elites. Alternatively, popular role models may inspire individual behavior and preferences (Porter and Serra, 2020).

Our theoretical framework suggests that the three proposed social-identity-based interventions may work to reduce misinformation uptake through their impacts on reducing affective polarization. As a result, per our pre-analysis plan, we will analyze the effects of the interventions on the intermediate outcome of polarization as well as on the main outcome of interest, misinformation uptake.[7]

2.3 Main Effects

The theoretical overview above generates the following preregistered hypotheses:

H 1 A digital literacy intervention (Capacity) providing information about what misinformation is and how it polarizes society will not reduce polarization or the propensity to believe and disseminate misinformation.

H 2 An empathy intervention (Motivation) providing individuals with a narrative about the outgroup that elicits empathy will decrease affective polarization and through it, reduce the propensity to believe and disseminate misinformation.

H 3 A social norms intervention (Motivation) providing individuals with the perception of a diversity of experiences among the ingroup toward the outgroup will decrease affective polarization and

[7] Our pre-analysis plan is available at `https://osf.io/n28wr`.

through it, reduce the propensity to believe and disseminate misinformation.

H 4 A popularity intervention that demonstrates popularity can be achieved through positivity (Motivation) will decrease affective polarization and through it, reduce the propensity to believe and disseminate misinformation.

Although we do not have ex ante expectations about which of H2, H3, or H4 will generate the strongest effect, we expect that – because H1 will have no effect – H2, H3, and H4 will generate effects that are greater than that of H1 or than the pure control.[8]

3 Polarization in Côte d'Ivoire

Côte d'Ivoire has long been a polarized society. Two civil wars pitted Southerners against Northerners in a battle to define the true Ivoirian identity (Ivoirité) (Konate, 2004). Figure 1 illustrates the geographical divide between the North and the South, demarcated here by a buffer zone created after the First Civil War.

This regional cleavage has a long history. Relative prosperity in the 20th century led to in-migration from neighboring, mostly Muslim countries. The share of Muslims in Côte d'Ivoire increased from 6% in 1922 (who were mostly in the North) to 39% by the end of the century. In the most recent census from 2014, Christians constitute about 34% of the population compared to 42% Muslims. The country's first post-independent president, Félix Houphouët-Boigny, managed to maintain peaceful relations among the country's increasingly diverse population, owing in large part to the nation's long-lasting economic prosperity. But economic crises beginning in the 1980s followed by Houphouët-Boigny's death in 1993 brought identitarian tensions to the surface.

Two members of Houphouët-Boigny's government vied to take his place – Henri Bédié who was the former President of the National Assembly, and Alassane Ouattara who was the former Prime Minister. Bédié won and ruled until 1999 when he was overthrown in a coup. The catalyst for

[8] While the wording of hypotheses H1-H4 indicates an expected mediated effect, our main equation for analysis in our PAP indicated that we would estimate the main effect of each intervention on misinformation uptake.

Figure 1: North-South Divide in Côte d'Ivoire

Note: Buffer Zone created after the end of the First Civil War in Spring 2007.

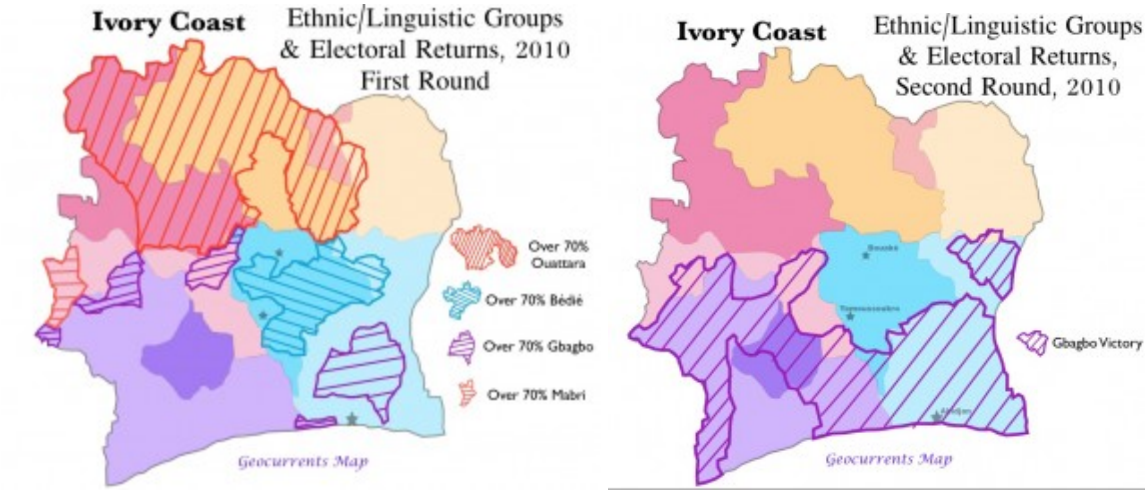
the first violent conflict that started shortly thereafter was the passage of a law requiring both parents of a presidential candidate to be born in Côte d'Ivoire. The law was seen as targeting the exclusion of Alassane Ouattara who was planning to stand in the 2000 election to replace the recently deposed Bédié.[9] Instead, Laurent Gbagbo was elected as the only main opposition candidate able to run in the 2000 presidential elections against the head of the transitional military government. Ouattara then won the next presidential election in 2010, postponed from 2005 due to the ongoing conflict.

While violent conflict in Côte d'Ivoire has mostly subsided, these deep regional divides are now enshrined in electoral politics as illustrated by Figure 2. Political polarization is exacerbated by overlapping identities. Southerners tend to be Christian, ethnic Akans (blue), and supporters of Gbagbo[10] and his now-opposition party FPI. Northerners tend to be Muslim, ethnic Mandinké or Burkinabé (pink), and supporters of Ouattara's now-ruling RHDP party. Supporters of the party of Houphouët-Boigny, PDCI, now led by Bédié are concentrated in the center of the country and less

[9]Ouattara had provided documents proving his and his parents' Ivoirian nationality, but those were called into question and annulled by the Bédié government.

[10]Gbagbo is ethnically Beté which is represented in dark purple, but he is also Christian like most of the South.

Figure 2: Overlapping Ethnic and Political Cleavages



Source: Geocurrents and Electoral Geography 2.0.

clearly fit into this North/South divide. At times, they have aligned with Ouattara – supporting him in the 2015 presidential elections, but they have also opposed him – running Bédié as a presidential candidate in 2020.

4 Research Design

Our empirical strategy was to develop interventions in partnership with an international non-governmental organization working on online misinformation and polarization on the ground, and to test their effectiveness via a two-wave panel survey experiment among Ivoirian youth in the country's economic capital, Abidjan. In this section, we describe these interventions, our sampling strategy, and our measurement strategy. We also describe our implementation partner and strategy.

4.1 Interventions and Random Assignment

We describe four interventions – one intervention meant to address the capacity constraint and three interventions meant to address the motivation constraint – that allow us to test hypotheses H1 through H4. The Motivation Interventions are all audio clips that are about five minutes in length

whereas the Capacity Intervention is a video clip that is about four minutes long.

Capacity Intervention:

Intervention 1 Digital literacy: our field partner had already developed and diffused content to inform young adults about online misinformation. A combination of two of the most informative videos were used as our digital literacy intervention.

Motivation Interventions:

Intervention 2 Empathy: A script, read by an actor[11], described a life challenge faced by a member of the outgroup. Outgroup identity was conveyed via the surname and the village of origin of the reader and his or her partner. Participants randomly assigned to this condition heard the script for the member of the *outgroup*. The gender of the narrator matched the gender of the respondent.

Intervention 3 Norms: Two focus group conversations (one focus group of 5-6 members of the Northern identity group; one focus group of 5-6 members of the Southern identity group), facilitated by professional enumerators, presented a discussion of people's everyday positive experiences and interactions with members of the outgroup. Participants randomly assigned to this condition heard excerpts of this conversation among members of their *ingroup*.

Intervention 4 Popularity: A famous comedian with a heavy social media presence delivered a message about being a positive influencer, and how altruism and positivity have been at the root of his success. We chose a comedian who purposefully does not self-identify with either group.

Assignment to the Capacity Intervention (vs. a placebo) was independent from assignment to one of the Motivation Interventions (vs. Control) per the below factorial design. The financial literacy placebo video, also produced by the local partner, was similar length to the digital literacy clip (about four minutes) and informed participants about how to set up a bank account and advice on

[11]To reduce additional heterogeneity, actors with neutral voices were chosen to read both versions of the script.

saving money. The control condition was simply the absence of any of the Motivation Interventions. Assignment probabilities are indicated in Table 1.[12]

Table 1: Factorial Design with Assignment Probabilities

	Control	Empathy	Norms	Popularity
Digital literacy	(A) $\frac{1}{5}$	(B) $\frac{1}{10}$	(C) $\frac{1}{10}$	(D) $\frac{1}{10}$
Placebo	(E) $\frac{1}{5}$	(F) $\frac{1}{10}$	(G) $\frac{1}{10}$	(H) $\frac{1}{10}$

Respondents were assigned with unequal probability to one of the eight cells of the factorial treatment conditions to maximize power to test our main hypotheses (see Appendix B for our power calculations). Random assignment was stratified on individual group status (Northern or Southern identity), enumerator group status (Northern or Southern identity), and on commune. The intuition for stratifying on enumerator group status is the research showing that enumerator ethnicity, and in particular the ethnic match between enumerator and respondent, might shape the answers the respondent provides (Adida et al., 2016). The intuition for stratifying on commune is that we intentionally sampled communes with different distributions of Northern and Southern residents in expectation of potentially heterogeneous effects conditional on this neighborhood characteristic. We discuss how we measure the main outcome variables in Section 4.3.

4.2 Research sites and study populations

We partnered with FieldPro research, a survey marketing organization in Côte d'Ivoire, to recruit young Ivoirians (18-30 year olds) in the country's urban center, Abidjan, to enroll them in an online program evaluation. While not the state capital, Abidjan is the country's urban and administrative center, a city of close to 5 million people, or one fifth of the country's population. Participation occurred in two waves (see Appendix A for the full timeline). In the first wave, FieldPro recruited participants using a random-walk face-to-face methodology in three of Abidjan's 10 communes.

[12] Appendix Table 5 presents the sample size in each cell at baseline and endline.

Because youth in this context tend not to be home during the day, enumerators were instructed to visit work sites, school sites, and other public locations where youth may be spending time during the day.

The requirements for participation were access to the internet, age (between 18 and 30 years old) and being born in Côte d'Ivoire. The reason for the latter requirement is that the study is interested in political effects and non-nationals may not identify with local parties and politics in the same way as nationals. Furthermore, the survey forces the respondent to make a choice between holding an identity as a Northerner or a Southerner in order to populate subsequent questions about ingroups and outgroups; these groupings make less sense for non-nationals.[13]

Abidjan's neighborhoods vary in their ethno-religious composition. Since we expected local-level ethno-religious heterogeneity to affect baseline polarization, we aimed to maximize variation on this construct in our sample communes. Together with the local survey team, we identified three communes that are similar in their economic status (less developed) but varied in terms of group make-up. We chose Abobo as a majority-Northerner commune, Port-Bouët as a majority-Southerner commune, and Yopougon as a mixed-group commune.

In the first wave, enumerators administered a questionnaire as well as the audio and/or video treatments. At the end of an initial battery of largely demographic questions, enumerators handed their tablet and headphones to the participant for self-administration of the video and audio intervention(s) per the above factorial design. Immediately following the intervention, respondents then self-administered a short battery of questions constituting a manipulation check and measurement of intermediate outcomes.

In the second wave, which occurred two to six weeks later, participants were sent a request to complete an online survey through their preferred mode of contact, i.e., Whatsapp, email, or SMS. Of the 2919 respondents recruited into the first wave of the study, 1891 consented to take the second

[13]Nationality is a sensitive subject in Côte d'Ivoire, and as such, enumerator training discussed issues that might arise around this qualification and how to address them.

survey. The attrition rate was thus slightly higher than the 30% anticipated by the local survey firm, even after extensive phone and in-person follow-up. Attrition is not, however, correlated with treatment status as depicted in Appendix Table 6. Participation in both surveys was remunerated with mobile transfers in the amount of 1500 CFA (about 3 USD) for baseline and 2000 CFA (about 4 USD) for endline.[14]

4.3 Outcomes of interest

In this section, we explain how we measure the key outcome variables proposed in our hypotheses. Our main outcomes are the correct identification of misinformation and intent to share misinformation. We measure affective polarization to test its role as both mediator and moderator, per our pre-analysis plan.

To measure the *Correct Identification of Information* and proclivity to *Share Misinformation*, we follow Badrinathan (2021) and Guess et al. (2020) and provide participants with a series of 12 news stories that vary in whether they are pro-Northern or pro-Southern (an equivalent number of each). To increase the chance of observing a reduction in belief in false stories, we include more false stories than true ones, eight and four, respectively.[15] Following the visual presentation of each news item, we ask the respondent two questions: 1) Do you believe this news story is false?, 2) How certain are you?, and 3) Would you share this news story with your friends or family? Since these questions may influence each other, we randomly assigned the order of the questions as well as the order of the news stories.

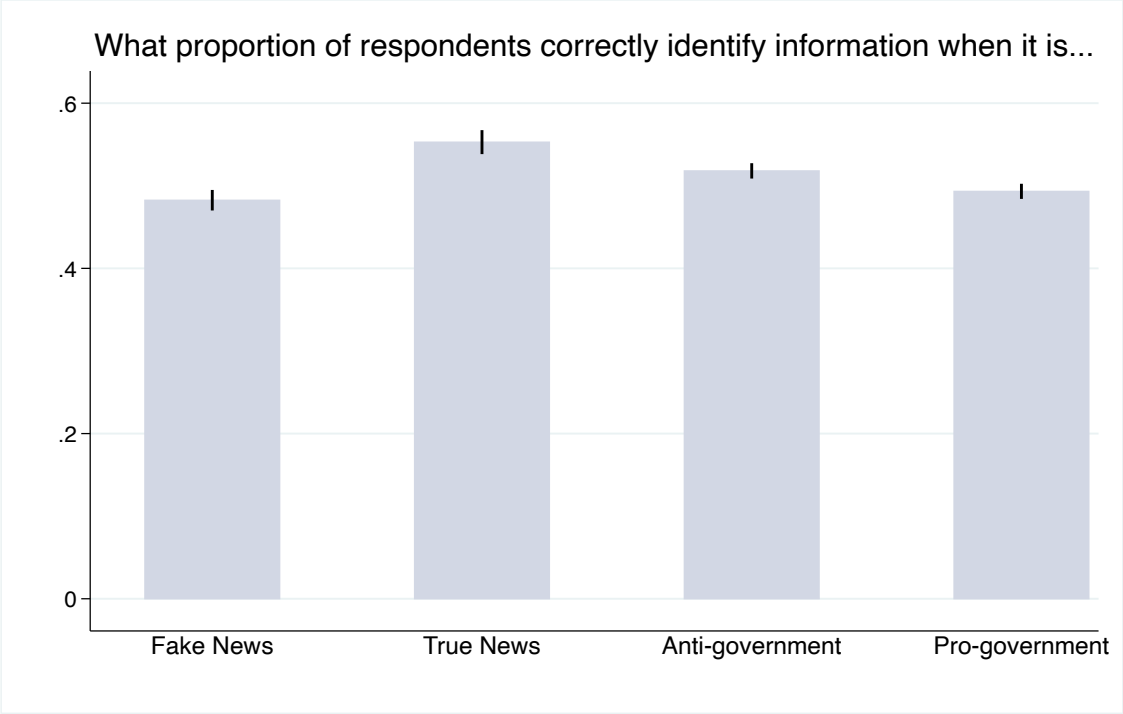
Our main outcome of interest is the proportion of the 12 news items that the respondent identified correctly, e.g., true news items are believed to be true and false news items are believed to be false. Figure 3 shows that respondents were on average fairly unsuccessful at identifying information with a rate about equivalent to merely guessing.[16] They are better at identifying

[14]We discuss our justification for this and other ethical considerations on Appendix E.

[15]Guay et al. (2022) propose that an appropriate test of interventions to reduce misinformation should include true and false content, and an analysis of the intervention's effects on the identification of both true and false content.

[16]In the discussion section, we disaggregate the outcome measure by a subjectively coded measure of difficulty to

Figure 3: Describing Correct Identification of Information



true news compared to false news, and they are somewhat better at identifying anti-government information relative to pro-government information. Correctly identifying news is positively correlated with being a woman, being older, identifying as a Northerner, and residing in Abobo (the most Northerner commune).

To measure *Affective polarization*, we use a battery of questions combining measures widely used in the literature. Specifically, we include a feeling thermometer for the ingroup and outgroup and trust questions for both groups; for these, we assess the difference between ingroup and outgroup responses. We also include a question on threat perception. We create an additive index of these component variables as our measure of affective polarization.

4.4 Estimation of Treatment Effects

To estimate treatment effects, we use the following equation:

assess whether treatment effects are conditional on difficulty of the information task.

$$Y_i = \alpha + \beta_1 Capacity_i + \beta_2 Motivation_i + X_i'\Gamma + \varepsilon_i \quad (1)$$

where Y_i is the relevant outcome measure, α is a constant term, β_1 is the average treatment effect of the Capacity Intervention (digital literacy relative to placebo), β_2 indicates the treatment effect of each of the Motivation Interventions (Norms, Empathy and Popularity) relative to Control, and X_i' is a vector of the blocking variables.

5 Results

Before turning to the estimation of treatment effects, we first consider whether respondents were more likely to believe information was true (false) when it was aligned (unaligned) with their identity. If respondents are indeed engaging in motivated reasoning, then those identifying as Southerners should be more likely to identify anti-government news as true (as there is currently a president from the North). Indeed, Southerners are 7.4 percentage points more likely to identify anti-government news items as true. Since we only have two groups and a binary outcome, this also implies that Northerners are more likely to identify anti-government news items as false.

We do not, however, observe a pattern of motivated reasoning with respect to pro-government news. While motivated reasoning would suggest that Northerners should be more likely to identify pro-government information as true (and Southerners more likely to identify pro-government news as false), there is no difference between the identity groups here. This could be driven by a ceiling effect as pro-government news is judged to be true 55% of the time relative to anti-government news which is believed to be true only 50% of the time. Alternatively, it could reveal something about the nature of motivated reasoning – that it is more likely to occur with anti-government than pro-government news. This could be, e.g., because there is more uncertainty around the verity of anti-government than pro-government news.

5.1 Correct Identification of Information

To test the effects of treatment on the *Correct Identification of Information*, we estimate Equation 1.[17] Figure 4 produces coefficient plots for the main outcome variable that simply averages over an indicator for whether the respondent correctly identified each of the 12 news items as true or false. Consistent with H 1, the Digital Literacy intervention has no effect on correctly identifying information. By contrast, we see effects of two of the three Motivation interventions, but not always in the expected direction. Consistent with H 2, the Empathy intervention increases the respondent's ability to correctly identify information.[18] But in contrast to H 3, the Norms intervention *decreases* the respondent's ability to correctly identify information. The Popularity intervention has no effect.

Yet this simple measurement strategy, which we had pre-registered, fails to adequately identify motivated reasoning, because Northerners and Southerners are differently motivated to accept or reject pro and anti-government information. Indeed, a Northerner engaging in motivated reasoning should be motivated to incorrectly identify pro-government fake news as true. In this case, we would expect the Motivation treatment to have a *positive* effect of correct identification of information; the treatment would be reducing the motivation to answer incorrectly. However, a Northerner receiving anti-government fake news should be more likely to correctly identify it as fake relative to a Southerner who might be motivated to believe the anti-government news. In this case, we would expect the Motivation treatment to have a *negative* effect on the correct identification of information; the treatment would be reducing the motivation to answer correctly.[19]

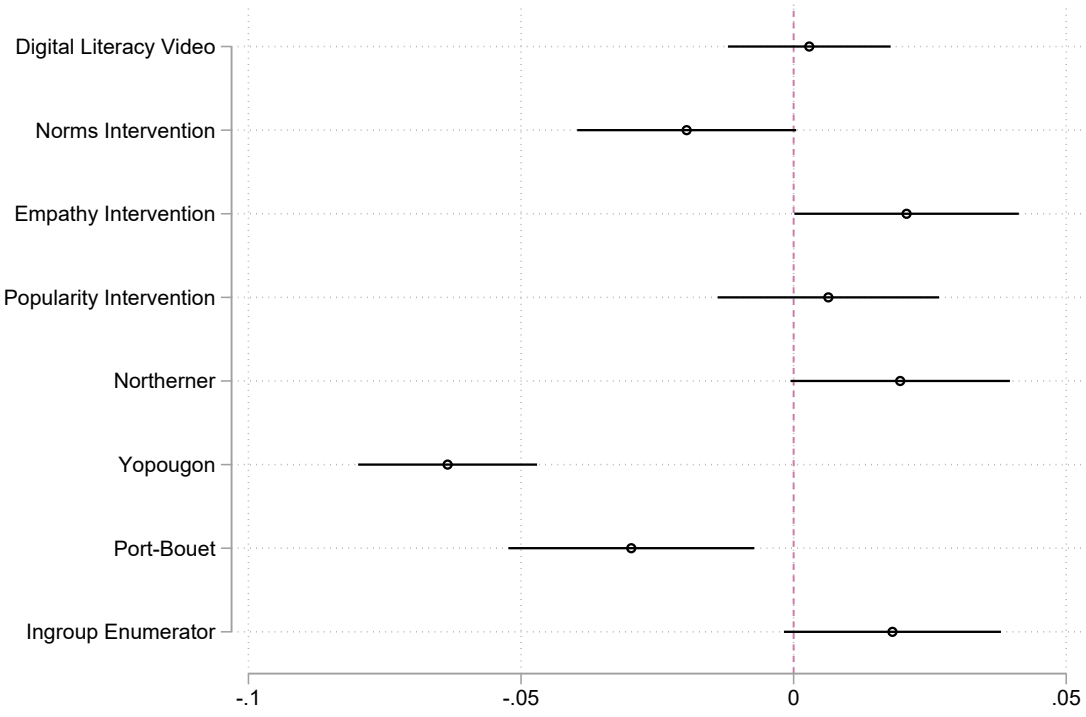
By collapsing news items that the respondent is motivated to both answer correctly and incor-

[17] These analyses drop the 336 respondents who self-reported opposing identities at baseline and endline. If the true identity was different than the one given at baseline, then the Motivation interventions would work in the opposite way as intended. Appendix C illustrates that our two main findings are robust to using the full sample of participants. It also describes an exercise to elicit true identities from an independent third party and test the robustness of findings on the sample for which self-reported baseline identities match the independent report but not the endline report. The findings are substantively similar in these additional models, although some treatment effects are no longer statistically significant at conventional levels.

[18] This is consistent with results in (Bowles et al., 2022), who find that an empathetic podcast was one of the most effective methods for improving misinformation identification.

[19] We thank Andrew Little for this insight.

Figure 4: Average Treatment Effects on Correct Identification of Information



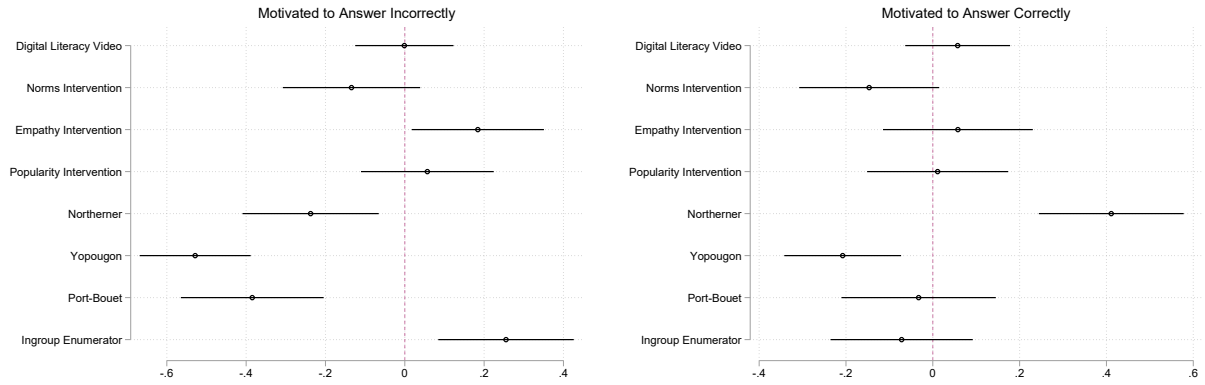
Note: This coefficient plots estimates Equation 1 on our pre-specified measure of correct identification of information, representing the proportion of 12 news items that the respondent correctly identified as true or false. For this and all subsequent coefficient plots, the omitted group for the commune indicators is Abobo. Complete results are presented in Column 1 of Appendix Table 7.

rectly, our pre-specified coding of the outcome variable may be masking important heterogeneity. To address this, we create secondary outcome variables that allow us to observe differential effects of treatment depending on whether the respondent was motivated to answer correctly or incorrectly. We create two count variables: *Correct Motivated* counts the number of news items the respondent correctly identified when they were motivated to do so; *Correct Unmotivated* counts the number of news items the respondent correctly identified when they were unmotivated to do so. Because the news items were equally distributed between pro- and anti-government slant, these count variables range from 0 to 6.

To test the idea that the Motivation treatment increases correct identification of information when people are motivated to get the answer wrong, we estimate Equation 1 for the the new outcome variable Correct Unmotivated (left panel in Figure 5). To test the idea that the Motivation treatment decreases correct identification of information when people are motivated to get the answer right, we estimate Equation 1 for the new outcome variable Correct Motivated (right panel in Figure 5). Here, the findings are consistent with our expectations. The positive effect of the Empathy treatment is coming from the set of news items that the respondent was motivated to identify incorrectly. The negative effect of the Norms treatment is apparent in both cases but slightly higher in the case where respondents were already motivated to correctly identify news items ($p<0.10$).

We note that this alternative outcome measurement strategy, while not pre-registered, is the only one that can adequately identify motivated reasoning. Indeed, if we measure only the individual's likelihood of correctly identifying information, then we cannot distinguish between an individual who is motivated to identify information correctly and an individual who has strong priors about the information's accuracy. It is only by differentiating between the motivation to answer incorrectly and the motivation to answer correctly that we can assert that we have captured motivated reasoning.

Figure 5: Average Treatment Effects on Correct Identification of Information



Note: These coefficient plots estimates Equation 1 on our secondary outcome measures of correct identification of information when respondents are motivated to answer incorrectly (left panel) and correctly (right panel). Complete results are presented in Columns 2 and 3 of Appendix Table 7 for the left and right panels, respectively.

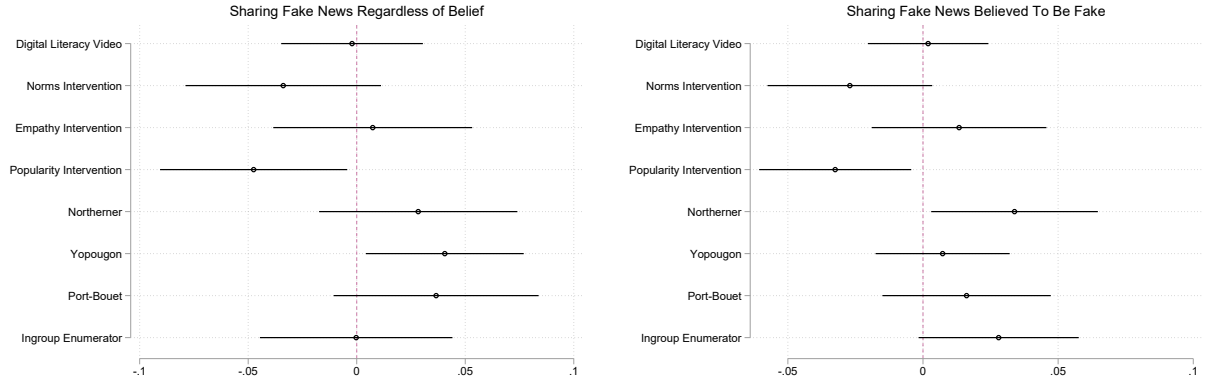
5.2 Sharing Misinformation

To test the effects of treatment on *Sharing Misinformation*, we again construct two measures (we did not specify exactly how we would measure this construct in our pre-analysis plan). The first is the most straightforward: we subset the 12 news items to the eight that were fake news. We then estimate the proportion of those items that respondents said they would share with others. The mean of this variable is 0.40.

However, some respondents believed that the information they said they would share was true and so would unknowingly be sharing misinformation. We thus create a second outcome variable that again subsets to the eight fake news items. But this time, we only count the proportion of items the respondent says they would share *and* also believes to be false. So this outcome captures whether the respondent would knowingly share misinformation. The mean of this variable is 0.18.

Figure 6 plots the coefficients of Equation 1 estimated on both versions of the outcome variable. The patterns are similar across both: the coefficients on the Digital Literacy and Empathy interventions are close to zero whereas the coefficients on the Norms and Popularity interventions

Figure 6: Average Treatment Effects on Sharing Misinformation



Note: These coefficient plots estimate Equation 1 on two measures of sharing of misinformation. The left panel represents the proportion of the eight fake news items that the respondent says they would share. The right panel represents the proportion of eight fake news items that the respondent both knows to be fake news and would also share. Complete results are presented in Columns 4 and 5 of Appendix Table 7 for the left and right panels, respectively.

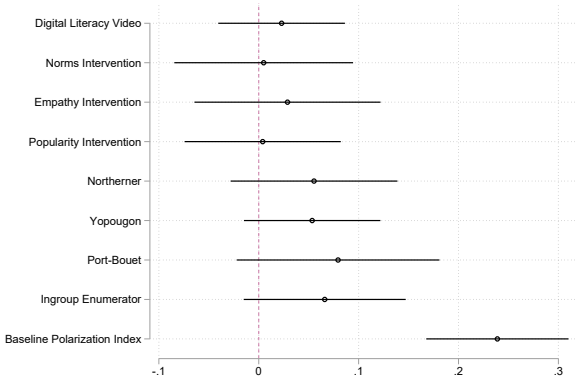
are negative and larger (0.12 to 0.15 standard deviations).[20] The negative effect on sharing of misinformation is consistent with H 4 (Popularity), although this effect was expected to obtain for the Norms and Empathy intervention as well. The coefficient on the Norms treatment indicator is not statistically significant at conventional levels; $p = 0.14$ in the left panel and $p = 0.08$ in the right panel.

Following Guay et al. (2022), we analyze discernment between the sharing of true and false news. They find research designs most convincing when treatment reduces sharing of false information but not true information. This is the pattern we see in our data with respect to the Popularity intervention. As shown in Appendix Table 8, the Popularity intervention has a negative and statistically significant effect on sharing information believed to be false and correctly believed to be false (Columns 1 and 2), and no effect of sharing information believed to be true and correctly believed to be true (Columns 3 and 4).[21] Interestingly, the Norms intervention has a negative effect on sharing of both types of information (statistically significant at $p < 0.1$). Here, we do not

[20] Another way of conceiving of the outcome variable is counting any news item (true or false) that the respondent believes is false and then shares. We see a similar pattern of results for this test in Column 1 of Appendix Table 8.

[21] Column 5 shows that the intervention also has a net negative effect of sharing of any information, but the other columns clearly show that the effect is being driven by a reduction in sharing false information.

Figure 7: Average Treatment Effects on Affective Polarization



Note: These coefficient plots estimate Equation 1 on our endline measure of affective polarization, controlling for the baseline measure of affective polarization. Complete results are presented in Column 1 of Appendix Table 9.

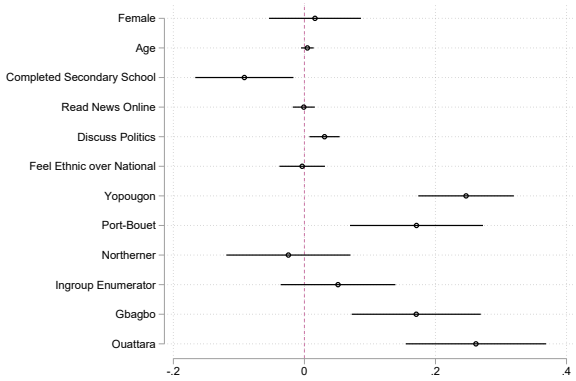
conclude that the effect of the Norms intervention on sharing of misinformation had the intended effect.

5.3 Polarization

We turn to an analysis of treatment effects on *Affective Polarization*, which we hypothesized would mediate the effect of treatment on the information outcomes. Figure 7 plots the coefficients for Equation 1 estimated on the affective polarization index described in Section 4. Because we also measured this index at baseline prior to the administration of treatment, we control for the baseline level of the outcome variable to increase precision. While we did not expect the Digital Literacy intervention to impact polarization, we did expect the three motivation interventions to have an effect. They clearly do not. Because we do not see statistically significant effects of treatment on our proposed mediator, affective polarization, we do not run the mediation analyses we proposed in the pre-analysis plan.

One explanation for this null finding is that the index is not doing a good job of capturing variation in this population. However, as we see in Figure 8, there are clear and sensible correlates of affective polarization at baseline. Supporters of the two main political rivals in Côte d'Ivoire

Figure 8: Correlates of Baseline Affective Polarization



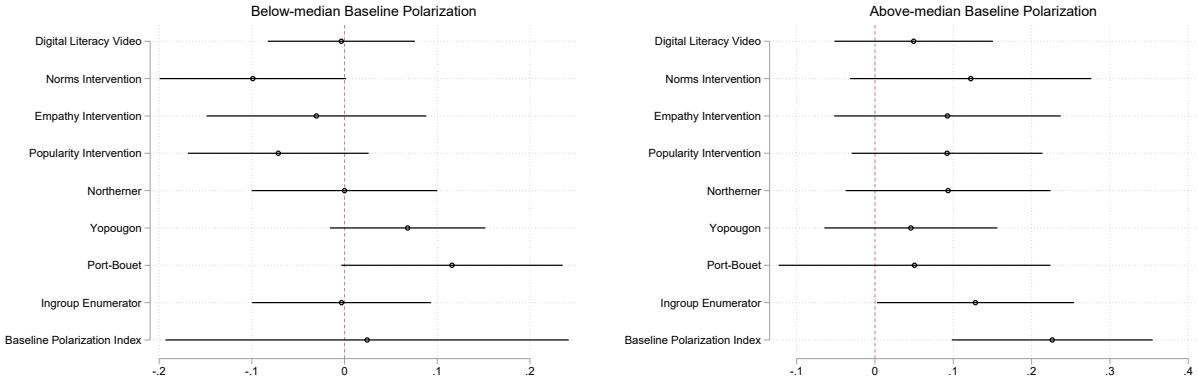
Note: These coefficient plots estimate Equation 1 on our endline measure of affective polarization, controlling for the baseline measure of affective polarization. The same specification is run on two subsets of our sample – below- and above-median baseline polarization – in the left and right panels, respectively. Complete results are presented in Columns 2 and 3 of Appendix Table 9.

are among the most polarized, as we would expect. Polarization is also positively correlated with being a Northerner, living in a more densely Northerner commune, and being more likely to discuss politics. It is negatively correlated with higher levels of education.

Alternatively, the null effect could be masking some heterogeneity in the population. We might expect the treatments to work differently on respondents with different baseline levels of affective polarization.[22] Figure 9 plots the effects of treatment on endline levels of affective polarization subsetting the sample by polarization levels at baseline (above- and below-median). Here, we see some evidence that the Motivation interventions may have been working differently among the more and less polarized respondents. In fact, it appears that the treatments may be further polarizing in that they decrease polarization among the least polarized and increase polarization among the most polarized. This trend is not statistically significant at conventional levels, but is most apparent (significant at 10%) for the Norms intervention.

[22]We pre-registered our expectation of heterogeneous treatment effects by baseline affective polarization, but did not specify a direction.

Figure 9: Average Treatment Effects on Affective Polarization by Baseline Polarization



6 Discussion

To summarize the results, the Digital Literacy intervention elicited no effect on the main outcomes of interest whereas each of the three Motivation interventions – Empathy, Norms, and Popularity – had a statistically significant effect on belief in or sharing of misinformation. This pattern of findings provides support for the idea that motivated reasoning, or the influence of one's identity on information processing, is a real constraint to efforts to reduce the ill effects of misinformation. This conclusion is consistent with the pattern of findings in Badrinathan's (2021) study wherein a digital literacy campaign worked among non-copartisans of the incumbent but not among co-partisans. She argues that because incumbent co-partisans exhibit greater attachment to the incumbent party relative to opposition partisans who are more fractionalized and less unified around a common identity, co-partisans would be more subject to motivated reasoning, limiting the effects of a digital literacy campaign.

The differential effects of each motivation intervention on our outcomes generates interesting findings. The more individualistic intervention (Empathy), which provides individuals the perspective of an outgroup experiencing a common life tragedy, has the expected positive effect on the more individualistic outcome – correct identification of information. By contrast, the more collective intervention (Popularity), which provides individuals with a message on the role of kindness

on being popular, has the expected negative effect on the more collective outcome – sharing of misinformation. While this interpretation is post hoc, it merits further investigation.

The non-significant effect of the digital literacy intervention accords with some of the mixed findings in the literature about the potential to affect information processing with short-term capacity interventions. While the intervention was taken directly from a set of digital literacy programming being promoted in Côte d'Ivoire, it was a relatively light-touch intervention compared to some of the others in the literature. For instance in Badrinathan's (2021) study in India, participants received an hour-long in-person media literacy training. This study's finding of no average effect of treatment was part of what motivated our expectation of a null result. However, the intervention we study was more intensive than that in Guess et al. (2020), which consisted of a text-based presentation of six to 10 tips about how to spot false news. While this latter study finds effects among a highly educated online sample in India, it finds no effects among a largely rural sample. Our sample of youth in Abidjan is probably more similar to the educated online sample in India and should thus add skepticism to the potential for digital literacy interventions to elicit positive impacts.

Given the low rate of correct identification, one might worry about the the difficulty of the information identification tasks used as the outcome variable. In the Guess et al. (2020) study, the base rate of correct identification of information among the online sample in India was also about 50% (and even lower in the rural sample). But the rate of correct identification in the Badrinathan (2021) study was much higher, at 83%, likely indicating a much easier task. To try to understand if the difficulty of the task is moderating the effect of the digital literacy treatment, we had two local youth – similar in profile to those in our sample – independently rate each of the 12 news items as easy, medium or hard to correctly identify. Averaging over the scores, they rated three items as easier, and six items as harder. We then run our main specification on the correct identification of information separately on the easier tasks and the harder tasks. Appendix Figure 11 provides suggestive evidence that the digital literacy intervention may have worked to increase correct identification of information among the easier tasks (p=0.16) but not the harder ones.

There is a remaining unexplained finding: the negative effect of the Norms intervention on the correct identification of information in the pre-registered analysis illustrated in Figure 4. This negative effect makes sense for the half of news items that individuals were motivated to answer correctly; the Norms intervention could have contributed to removing that motivation resulting in a reduction of correct responses. However, as illustrated in Figure 5, the negative effect looks substantively similar among the half of news items that participants were motivated to answer incorrectly and for which we expected a positive finding. Since the Norms intervention appeared to elicit differential effects on affective polarization based on initial levels of polarization, it could be that the negative effect of the Norms intervention on information processing is masking some heterogeneity as well. Interacting the treatment indicators with a binary indicator of above- and below-median baseline polarization, however, we find no evidence that the Norms intervention has any differential effect.

Another possible moderator of treatment is political sophistication. Walter et al. (2020) have shown that motivated reasoning is more prevalent among politically sophisticated individuals. But Vegetti and Mancosu (2020) find that political sophisticates are better able to tell real from false news. This poses an interesting paradox for our study: political sophisticates are more susceptible to motivated reasoning but less susceptible to fake news. Baseline affective polarization in our data is positively correlated with talking politics more frequently, consistent with the expectation in the literature. Also consistent with expectations from the literature, correct identification of information is positively correlated with talking politics more frequently. If we interact a binary indicator of above- and below-median frequency of talking politics with our treatments, we find that the negative effect of norms obtains only among the more "politically sophisticated" according to this measure. There was no differential effect of treatment on endline polarization. This pattern of findings could be consistent with the Norms intervention reducing the advantage that political sophisticates usually have in deciphering fake news, but this interpretation requires additional study.

Another way to diagnose where this effect is coming from is to observe which news items

are driving the result. We can divide the news items into four categories along the dimensions of true/false and pro/anti-government. The negative effect of the Norms intervention on the correct identification of information only appears for the anti-government fake news. This implies that the Norms intervention causes people to be overly generous about believing opposition news – they are more likely to think fake news favoring the opposition is true. If people perceive the incumbent party as being more likely to perpetuate fake news, then this reaction could be an over-correction. But this, too, is speculative and the unexpected finding merits further investigation.

7 Conclusion

This study adds to the small but growing literature that urges skepticism about the effectiveness of interventions to correct information in non-Western contexts (Badrinathan, 2021; Batista et al., 2022; Bowles et al., 2022). It is also the first to explicitly pair an information correction intervention with one that aims to reduce motivated reasoning. The results are enlightening: all three Motivation interventions shaped the proclivity of consumers to believe or share false information. Meanwhile, the Capacity intervention had zero effect on identifying or sharing misinformation. Our results have theoretical and empirical implications.

Theoretically, our study confirms the role that demand-based motivation factors play in perpetuating media bias and misinformation: consumers of information are motivated not just by accuracy goals, but also – and in polarized contexts more importantly – by directional goals: information that affirms one's social identity is more readily assimilated, whether or not it is true. As a result, the constraint on assimilating new information does not lie in an individual's *capacity* to assimilate new information, but rather in their *motivation* to assimilate or resist it.

Empirically, actors looking for strategies to reduce misinformation can use this theoretical insight to inform their interventions. Social scientific research on how social identity shapes individual behavior offers important insights into the role that social norms, empathy, and the quest

for self-esteem can play in shaping these directional goals. The finding on the empathy intervention, in particular, seems easily scalable: one simple perspective-getting narrative had a durable effect on individuals' rejection of misinformation they were motivated to believe. These types of narratives can be easily incorporated in any news story, and in fact they often are.[23]

Finally, while the unexpected effect of the Norms intervention raises additional questions, the fact that misinformation outcomes are sensitive to this type of manipulation of ingroup and outgroup norms is consistent with our other findings. We hope this evidence will encourage further innovation and testing around interventions aimed to reduce motivated reasoning in contexts of rampant misinformation.

[23] See, e.g., `https://www.npr.org/series/4516989/storycorps`.

References

Adida, Claire L., Karen E. Ferree, Daniel N. Posner and Amanda Lea Robinson. 2016. "Who,Äôs Asking? Interviewer Coethnicity Effects in African Survey Data." *Comparative Political Studies* 49(12):1630–1660.
URL: *https://doi.org/10.1177/0010414016633487*

Badrinathan, Sumitra. 2021. "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India." *American Political Science Review* pp. 1–17.

Baron, Hannah, Robert Blair, Donghyun Danny Choi, Laura Gamboa, Jessica Gottlieb, Amanda Lea Robinson, Steven Rosenzweig, Megan Turnbull and Emily A West. 2021. "Can Americans Depolarize? Assessing the Effects of Reciprocal Group Reflection on Partisan Polarization.".

Batista, Frederico, Natalia S Bueno, Felipe Nunes and Nara Pavão. 2022. "Fake News, Fact Checking, and Partisanship: The Resilience of Rumors in the 2018 Brazilian Elections.".

Boudreau, Cheryl. 2019. In *Oxford Handbook of Electoral Persuasion*, ed. Bernard Grofman, Elizabeth Suhay and Alexander H. Trechsel. Oxford University Press.

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, Shelley Liu and John Marshall. 2022. "Sustained exposure to fact-checks can innoculate citizens against misinformation in the Global South.".
URL: *https://www.dropbox.com/s/a6rnh3o7c97dcqn/WCW_Science_Submission.pdf?dl=0*

Broockman, David, Joshua Kalla and Sean Westwood. 2022. "Does affective polarization undermine democratic norms or accountability? Maybe not." *American Journal of Political Science* .

Bullock, J. 2020. Party cues. In *The Oxford Handbook of Electoral Persuasion*, ed. E. Suhay, B. Grofman and A.H. Trechsel. New York, NY: Oxford University Press.

Cialdini, Robert B., Stephanie L. Brown, Brian P. Lewis, Carole Luce and Steven L. Neuberg. 1997. "Reinterpreting the empathy-altruism relationship: when one into one equals oneness." *Interpersonal relations and group processes* 73(3).

Cook, John, Stephan Lewandowsky and Ullrich KH Ecker. 2017. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence." *PloS one* 12(5):e0175799.

Coppock, Alex. 2016. Positive, small, homogenous, and durable: political persuasion in response to information PhD thesis Columbia University.

Ehret, P. 2021. "Reaching Republicans on climate change." *Nature Climate Change* 11.

Fearon, J.D. and D.D. Laitin. 1996. "Explaining interethnic cooperation." *The American Political Science Review* 90(4).

Flynn, D.J., Brendan Nyhan and Jason Reifler. 2017. "The nature and origins of misperceptions: understanding false and unsupported beliefs about politics." *Advances in Political Psychology* 38(1).

Gatewood, Cooper, Alex Krasodomski-Jones, Zoe Fourel, Cecile Guerin, Charlotte Moeyens and Josh Smith. 2020. "Disinformation in the Ivory Coast, Case Study 2.".

Gentzkow, Matthew, Jesse M Shapiro and Daniel F Stone. 2015. Media bias in the marketplace: Theory. In *Handbook of media economics*. Vol. 1 Elsevier pp. 623–645.

Graham, Matthew H and Milan W Svolik. 2020. "Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States." *American Political Science Review* 114(2):392–409.

Guay, Brian, adam berinsky, Gordon Pennycook and David G Rand. 2022. "How To Think About Whether Misinformation Interventions Work.".
URL: *psyarxiv.com/gv8qx*

Guess, Andrew M and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. In *Social media and democracy: The state of the field, prospects for reform*, ed. Nathaniel Persily and Joshua A. Tucker. Cambridge: Cambridge University Press Cambridge pp. 10–33.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler and Neelanjan Sircar. 2020. "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* 117(27):15536–15545.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner and Jeremy M. Weinstein. 2007. "Why does ethnic diversity undermine public goods provision?" *The American Political Science Review* 101(4).

Hopkins, Daniel J., John Sides and Jack Citrin. 2019. "The Muted Consequences of Correct Information about Immigration." *The Journal of Politics* 81(1):315–320.
URL: *https://doi.org/10.1086/699914*

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1):1–17.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J. Westwood. 2019. "The Origins and Consequences of affective Polarization in the United States." *Annual Review of Political Science* 22:129–146.

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

Kahan, Dan M. 2016. The politically motivated reasoning paradigm, part I: what politically motivated reasoning is and how to measure it. In *Emerging trends in the social and behavioral sciences*, ed. Robert Scott and Stephen Kosslyn. John Wiley & Sons.

Kalla, Joshua and David Broockman. 2020. "Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting." *Preprint. Open Science Framework. https://doi. org/10.31219/osf. io/z2awt* .

Konate, Siendou A. 2004. "The politics of identity and violence in Côte d'Ivoire." *West African Review* 5.

Lelkes, Yphtach. 2016. "Mass Polarization: Manifestations and Measurements." *Public Opinion Quarterly* 80(1).

Levendusky, Matthew. 2018. "Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization?" *Journal of Politics* 80:59–70.

Lewandowsky, Stephan, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz and John Cook. 2012. "Misinformation and its correction: Continued influence and successful debiasing." *Psychological science in the public interest* 13(3):106–131.

Little, Andrew. 2021. "Detecting motivated reasoning." *OSF Preprints* .

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky and Ralph Hertwig. 2022. "Digital media and democracy: a systematic review of causal and correlational evidence worldwide." *Nature Human Behavior* .

Mason, Lilliana. 2018. "Losing Common Ground: Social Sorting and Polarization." *The Forum* 16(1):47–66.

McCauley, John F. 2017. *The logic of ethnic and religious conflict in Africa*. Cambridge University Press.

McClendon, Gwyneth. 2018. *Envy in Politics*. Princeton University Press.

McCoy, Jennifer, Tahmina Rahman and Murat Somer. 2018. "Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities." *American Behavioral Scientist* 62(1):16–42.

McGuire, William J. 1964. "Inducing resistance to persuasion. Some contemporary approaches." *CC Haaland and WO Kaelber (Eds.), Self and Society. An Anthology of Readings, Lexington, Mass.(Ginn Custom Publishing) 1981, pp. 192-230.* .

Nyhan, Brendan, Ethan Porter, Jason Reifler and Thomas J Wood. 2020. "Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability." *Political Behavior* 42(3):939–960.

Paluck, Elizabeth Levy and Donald P Green. 2009. "Prejudice reduction: What works? A review and assessment of research and practice." *Annual review of psychology* 60:339–367.

Paluck, Elizabeth Levy, Hana Shepherd and Peter M. Aronow. 2016. "Changing climates of conflict: A social network experiment in 56 schools." *Proceedings of the National Academy of Sciences* 113(3):566–571.
URL: *https://www.pnas.org/content/113/3/566*

Peterson, Erik and Shanto Iyengar. 2021. "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?" *American Journal of Political Science* 65(1):133–147.

Pink, Sophia L., James Chu, James N. Druckman, David G. Rand and Robb Willer. 2021. "Elite party cues increase vaccination intentions among Republicans." *Proceedings of the National Academy of Sciences* 118(32).
URL: *https://www.pnas.org/content/118/32/e2106559118*

Porter, Catherine and Danila Serra. 2020. "Gender differences in the choice of major: the importance of female role models." *The American Economic Review: Applied Economics* 12(3).

Porter, Ethan and Thomas J Wood. 2022. "Political Misinformation and Factual Corrections on the Facebook News Feed: Experimental Evidence." *Journal of Politics* .

Suhay, Elizabeth, Emily Bello-Pardo and Brianna Maurer. 2018. "The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments." *The International Journal of Press/Politics* 23(1).

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American journal of political science* 50(3):755–769.

Tajfel, H. and J.C. Turner. 1986. The Social Identity Theory of Intergroup Behavior. In *Psychology of Intergroup Relation*, ed. S. Worchel and W.G. Austin. Chicago: Hall Publishers pp. 7–24.

Traberg, Cecilie S, Jon Roozenbeek and Sander van der Linden. 2022. "Psychological inoculation against misinformation: Current evidence and future directions." *The ANNALS of the American Academy of Political and Social Science* 700(1):136–151.

Tucker, Joshua A., Yannis Theocharis, Margaret E. Roberts and Pablo Barbéra. 2017. "From liberation to turmoil: social media and democracy." *Journal of Democracy* 28(4).

Tucker, Joshua Aaron, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *SSRN* .
URL: *https://ssrn.com/abstract=3144139 or http://dx.doi.org/10.2139/ssrn.3144139*

Tully, Melissa, Emily K Vraga and Leticia Bode. 2020. "Designing and testing news literacy messages for social media." *Mass Communication and Society* 23(1):22–46.

Vegetti, Federico and Moreno Mancosu. 2020. "The impact of political sophistication and motivated reasoning on misinformation." *Political Communication* 37(5):678–695.

Voelkel, Jan G., Michael N. Stagnaro, James Chu, Sophia Pink, Joseph S. Mernyk, Chrystal Redekopp, Matthew Cashman, Qualifying Democracy Challenge Submitters, James N. Druckman, David G. Rand and Robb Willer. 2022. "Megastudy identifying successful interventions to strengthen Americans' democratic attitudes.".
URL: *https://www.strengtheningdemocracychallenge.org/paper*

Vraga, Emily K, Leticia Bode and Melissa Tully. 2020. "Creating news literacy messages to enhance expert corrections of misinformation on Twitter." *Communication Research* p. 0093650219898094.

Walter, Nathan, Jonathan Cohen, R Lance Holbert and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom." *Political Communication* 37(3):350–375.

White, Ismail K., Chryl N. Laird and Troy D. Allen. 2014. "Selling Out?: The Politics of Navigating Conflicts between Racial Group Interest and Self-interest." *American Political Science Review* 108(4):783,Äì800.

Zhuravskaya, Ekaterina, Maria Petrova and Ruben Enikolopov. 2020. "Political effects of the internet and social media." *Annual Review of Economics* 12:415–438.

Appendices

A Timeline

- August 23- October 22, 2021: Creation of intervention content; development of PAP; programming of questionnaires; IRB approval
- November 11-16, 2021: Training and piloting
- November 17-December 7, 2021: Recruitment and baseline survey administration in waves
- December 1-January 18, 2022: Online fielding of endline questionnaire

B Power Calculations

We estimated power on the basis of an endline sample size of 2000 participants. Our survey team anticipated a maximum rate of attrition of 30% between baseline and endline. We thus recruited about 2900 participants at baseline to yield approximately 2000 participants at endline.

Our power calculation assumes a conservative 0.2 standard deviation minimum detectable effect size. We base this off of the literature on affective polarization. There, an intensive in-person workshop reduced affective polarization by 0.3 standard deviations (Baron et al., 2021) whereas a less intensive treatment in which respondents were asked to read and reflect on a news article reduced affective polarization by 0.2 standard deviations (Levendusky, 2018). Our intervention likely falls somewhere in between these two in terms of intensity but closer to the news article prime, so we choose 0.2 to be conservative.

With a final sample size of 2000, an assumed minimum detectable effect size of 0.2 standard deviations, and a statistical significance level of 0.05, this design will allow be powered for our hypothesis tests as follows:

- For H1, which compares digital literacy to the placebo with N=400 in each group, the power of the test is 0.81.

- For H2-H4, which compares each of Interventions 2-4, respectively, to the Control condition with N=400 in each intervention group and N=800 in Control, the power of the test is 0.90.

C Robustness Checks

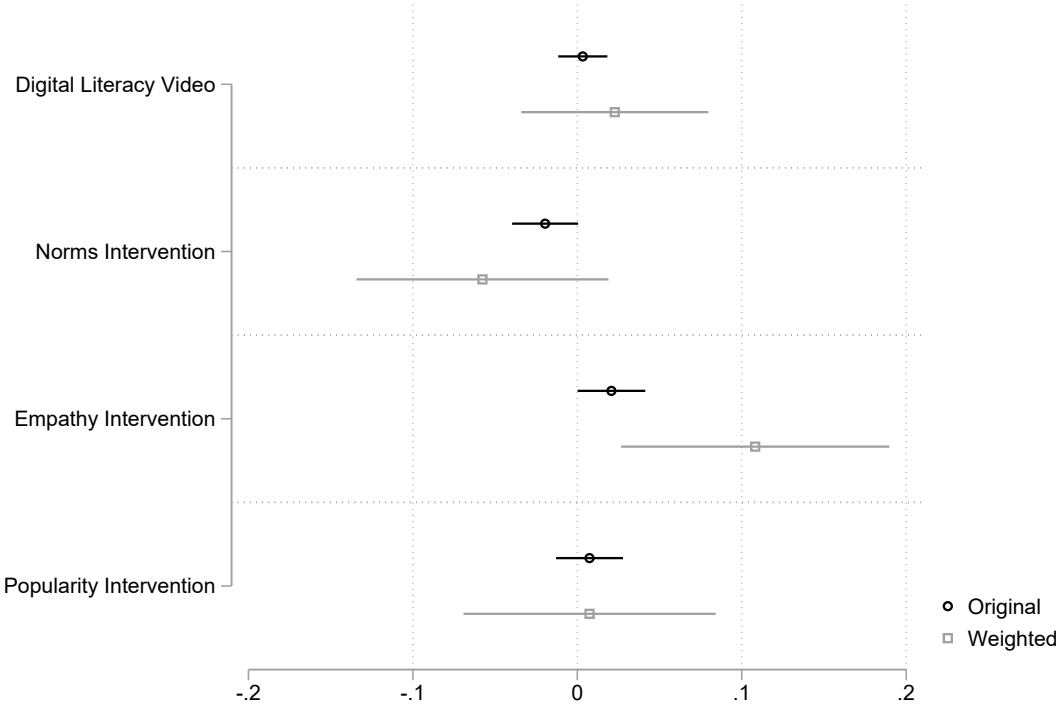
Here, we examine the robustness of our main results to two additional considerations. First, we examine how taking account of the certainty of respondents answers affects the estimation of treatment effects. Second, we examine whether our results are robust to different exclusion rules.

In addition to asking respondents a binary question about whether each news item was false, we also asked how certain they were about their answer. On average, respondents were fairly certain. On a 4-point scale from very uncertain to very certain, the mean response was 2.9 across all news items. People who are more likely to say they discuss politics at baseline are more certain about their answers as are people who share the group of the enumerator. There are no treatment effects on average level of certainty. However, if we weight each response by level of certainty, the negative effect of the Norms treatment is amplified as shown in Figure 10.

As we were verifying data quality at the start of the second survey, we found that a substantial portion of respondents (about 20 percent) were reporting a different identity (Northerner or Southerner) than the identity they reported in the first wave. We decided we would need to drop these respondents in the data analysis because we did not know the true identity of the respondent. If the identity was different than the one given at baseline, then the Motivation interventions would work in the opposite way as intended because the audio assigned to the respondent would be about the wrong group (assignment to treatment was based on self-reported identity at baseline). The analyses presented in the main text thus drop the 336 respondents that self-reported opposing identities at baseline and endline.

However, it is conceivable that treatment is correlated with this proclivity to change identity reporting which could bias estimates of treatment effects. Indeed, as shown in Table 2, the Empathy

Figure 10: Average Treatment Effects on Correct Identification of Information, Weighted by Certainty of Response



treatment makes respondents more likely to switch their identity between the two survey waves (statistically significant at $p < 0.1$). We thus check robustness of our results to using the full sample (keeping the people who switched identities between baseline and endline). We would expect our main results to attenuate somewhat given the inclusion of respondents who got the "wrong" treatment and would thus be expected to react in the opposite direction as the one hypothesized. Indeed, in Column 4 of Tables 3 and 4, we see the magnitude of the estimated effect of the Empathy and Popularity treatments on correct identification and sharing of misinformation, respectively, attenuate. But in each case, the coefficient is statistically significant at $p < 0.1$.

Table 2: Average Treatment Effects on Indicators of Mismatch for Exclusion from Analysis

	(1) Self-Reported	(2) Verified by Any	(3) Verified by All
Digital Literacy Video	0.007	0.015	0.010
	(0.018)	(0.013)	(0.012)
Norms Intervention	0.032	-0.024	-0.014
	(0.024)	(0.016)	(0.015)
Empathy Intervention	0.047^{+}	0.010	0.023
	(0.025)	(0.018)	(0.017)
Popularity Intervention	0.033	-0.000	0.008
	(0.024)	(0.018)	(0.016)
Northerner	0.135^{***}	0.156^{***}	0.136^{***}
	(0.025)	(0.020)	(0.018)
Yopougon	0.062^{**}	0.052^{***}	0.041^{**}
	(0.020)	(0.015)	(0.014)
Port-Bouet	0.029	0.035^{*}	0.030^{+}
	(0.025)	(0.018)	(0.016)
Ingroup Enumerator	-0.022	-0.036^{*}	-0.037^{*}
	(0.024)	(0.018)	(0.016)
Constant	0.091^{**}	0.029	0.024
	(0.029)	(0.021)	(0.019)
Observations	1832	1885	1885

OLS model with robust standard errors, blocking variables included.
$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Models indicate which respondents were excluded on the basis of their baseline identity not matching the identity assumed by treatment.

As another way to address this issue, we attempt to code the "true" identity of each of these

336 respondents. To do this, we had three Ivoirian individuals independently code the Northern or Southern identity of each respondent based on their name, religion, ethnicity, region of birth, and family's region of origin. We then create a new exclusion variable based on this independent coding. Among the 336 respondents with mismatched identities at baseline and endline, we only exclude those whose baseline identity could not be independently verified by a) any coder, or b) all three coders. This amounts to 172 and 144 respondents, respectively. When we do this, our exclusion variable is no longer correlated with treatment status as indicated in Columns 2 and 3 of Table 2.

In Tables 3 and 4, we present results for our main analyses of average treatment effects on identification and sharing of misinformation using each of these definitions for exclusion, in turn. The first model reproduces the estimates presented in the above coefficient plots which drop anybody who gives a different identity in the first and second wave of the survey. The second model only excludes respondents who gave a different identity and whose baseline identity could not be verified by any coder; the third model excludes respondents who gave a different identity and whose baseline identity could not be verified by all coders. The findings are substantively similar in these additional models although the estimated positive treatment effect of the Empathy intervention on correct identification of information is no longer statistically significant.

One reason we choose to prioritize the first specification in the main text is because the exercise of externally coding respondent identity introduces its own form of bias. Digging deeper into the data on identity switchers, we can infer that there are at least two reasons people might switch reported identities between baseline and endline. First, respondents might not be paying attention and so misreport at baseline or endline. If they misreport at baseline, they receive the wrong treatment. If they misreport at endline, it is likely that their responses are of poor quality because they are not paying close attention to the survey (the endline survey was online, so this is especially important). A second reason for switching is that respondents may actually be on the fence. While we did not anticipate this possibility because of the political and cultural salience of the Northern and Southern identities, it turns out that there are regions and ethnic groups in the center of the

country that have a history of being unaligned with either group. Attention-paying respondents may have thus honestly switched groups across surveys because of a weak attachment to either. Indeed, among those who say Northerner at baseline, those from the unaligned groups are more likely to switch identities at endline than those from the aligned groups (38% vs. 25%). We would be especially interested in excluding the first type of respondent from our survey but not the second. In fact, treatment may be even more effective among the second type of respondent because they are the least attached to their group identity.

The coders are only able to address one of these three problems. Coders will be better at identifying the "true" identity of more aligned groups. This means the coder specifications will exclude the people who gave the "wrong" identity at baseline and thus got the wrong version of the treatment. However, these specifications include respondents who gave the "wrong" identity at endline and thus were likely to be inattentive when answering the questions measuring our outcomes of interest. Also, coders are not very good at identifying the "true" identity of unaligned groups. Among the entire sample (not just the people answering differently at baseline/endline), 91% of those identifying as Southerners at baseline were identified as Southern by coders. By contrast, only 51% of those identifying as Northerner at baseline were coded as such. Looking even deeper, we can separate out – among those who said they were Northerners at baseline – people who are from more central regions that are likely to be unaligned with either group. Here, coders only validated the baseline identity as Northerner 14% of the time for the unaligned groups compared to 62% for others.

Therefore, the result of the coding exercise is that we are more likely to keep people in the sample who are inattentive at endline and drop people who are from less aligned identity groups and were honestly switching. This could help explain why we see the effect size moderate in Columns 2 and 3 of the robustness tests.

Table 3: Average Treatment Effects on Correct Identification of Information When Motivated to Answer Incorrectly

	(1) Self-Reported	(2) Verified by Any	(3) Verified by All	(4) Full Sample
Digital Literacy Video	-0.001	0.024	0.017	-0.009
	(0.063)	(0.062)	(0.061)	(0.059)
Norms Intervention	-0.134	-0.155^{+}	-0.153^{+}	-0.140^{+}
	(0.088)	(0.086)	(0.085)	(0.083)
Empathy Intervention	0.184*	0.117	0.117	0.138^{+}
	(0.085)	(0.084)	(0.083)	(0.079)
Popularity Intervention	0.057	-0.014	-0.007	0.023
	(0.085)	(0.084)	(0.083)	(0.080)
Northerner	-0.237**	-0.231**	-0.211*	-0.198*
	(0.088)	(0.085)	(0.084)	(0.079)
Yopougon	-0.528***	-0.448***	-0.438***	-0.394***
	(0.071)	(0.069)	(0.069)	(0.066)
Port-Bouet	-0.385***	-0.360***	-0.340***	-0.281**
	(0.092)	(0.091)	(0.090)	(0.087)
Ingroup Enumerator	0.255**	0.240**	0.241**	0.220**
	(0.087)	(0.084)	(0.083)	(0.079)
Constant	3.173***	3.112***	3.097***	3.079***
	(0.112)	(0.106)	(0.105)	(0.102)
Observations	1502	1713	1741	1885

OLS model with robust standard errors, blocking variables included.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Models indicate which respondents were excluded on the basis of their baseline identity not matching the identity assumed by treatment.

Table 4: Average Treatment Effects on Sharing of Information Known to Be False

	(1) Self-Reported	(2) Verified by Any	(3) Verified by All	(4) Full Sample
Digital Literacy Video	0.002	-0.002	-0.001	-0.006
	(0.011)	(0.011)	(0.011)	(0.011)
Norms Intervention	-0.027^{+}	-0.019	-0.020	-0.021
	(0.016)	(0.015)	(0.015)	(0.015)
Empathy Intervention	0.013	0.002	0.003	0.003
	(0.016)	(0.015)	(0.015)	(0.015)
Popularity Intervention	-0.032*	-0.032*	-0.032*	-0.023^{+}
	(0.014)	(0.014)	(0.014)	(0.014)
Northerner	0.034*	0.028^{+}	0.028^{+}	0.025^{+}
	(0.016)	(0.015)	(0.015)	(0.014)
Yopougon	0.007	0.008	0.008	0.006
	(0.013)	(0.012)	(0.012)	(0.012)
Port-Bouet	0.016	0.018	0.022	0.014
	(0.016)	(0.015)	(0.015)	(0.015)
Ingroup Enumerator	0.028^{+}	0.020	0.019	0.009
	(0.015)	(0.014)	(0.014)	(0.014)
Constant	0.148***	0.161***	0.160***	0.176***
	(0.018)	(0.017)	(0.017)	(0.017)
Observations	1500	1687	1713	1854

OLS model with robust standard errors, blocking variables included.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Models indicate which respondents were excluded on the basis of their baseline identity not matching the identity assumed by treatment.

D Additional Tables and Figures

In this section, we present additional tables and figures referenced in our main paper.

Table 5: Details on Sample Size at Baseline and Endline

	Control	Empathy	Norms	Popularity
Digital literacy	**577**	**288**	**287**	**293**
	366	181	190	187
Placebo	**606**	**278**	**296**	**293**
	397	180	186	204

Sample size at baseline in **bold** font; at endline in regular font.

Table 6: Average Treatment Effects on Attrition in Follow-up Survey

	(1)
Digital Literacy Video	0.016
	(0.017)
Norms Intervention	-0.001
	(0.023)
Empathy Intervention	0.003
	(0.023)
Popularity Intervention	-0.026
	(0.023)
Northerner	0.046*
	(0.023)
Yopougon	-0.137***
	(0.019)
Port-Bouet	0.211***
	(0.023)
Coethnic Enumerator	0.026
	(0.023)
Constant	0.313***
	(0.028)
Observations	2910

OLS model with robust standard errors, blocking variables included. $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 7: Average Treatment Effects on Misinformation Outcomes

	(1) Correct Identification of Information	(2) Correct Unmotivated	(3) Correct Motivated	(4) Sharing Fake News Regardless of Belief	(5) Sharing Fake News Believed to Be Fake
Digital Literacy Video	0.003	-0.001	0.057	-0.002	0.002
	(0.008)	(0.063)	(0.062)	(0.017)	(0.011)
Norms Intervention	-0.020^{+}	-0.134	-0.147^{+}	-0.034	-0.027^{+}
	(0.010)	(0.088)	(0.082)	(0.023)	(0.016)
Empathy Intervention	0.021*	0.184*	0.058	0.007	0.013
	(0.010)	(0.085)	(0.088)	(0.023)	(0.016)
Popularity Intervention	0.006	0.057	0.011	-0.047*	-0.032*
	(0.010)	(0.085)	(0.083)	(0.022)	(0.014)
Northerner	0.020^{+}	-0.237**	0.411***	0.028	0.034*
	(0.010)	(0.088)	(0.085)	(0.023)	(0.016)
Yopougon	-0.063***	-0.528***	-0.208**	0.041*	0.007
	(0.008)	(0.071)	(0.069)	(0.019)	(0.013)
Port-Bouet	-0.030**	-0.385***	-0.033	0.037	0.016
	(0.011)	(0.092)	(0.091)	(0.024)	(0.016)
Ingroup Enumerator	0.018^{+}	0.255**	-0.072	-0.000	0.028^{+}
	(0.010)	(0.087)	(0.084)	(0.023)	(0.015)
Constant	0.518***	3.173***	3.027***	0.371***	0.148***
	(0.013)	(0.112)	(0.108)	(0.028)	(0.018)
Observations	1499	1502	1502	1500	1500

OLS model with robust standard errors, blocking variables included.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8: Average Treatment Effects on Sharing of Information by Information Type

	(1) Think False	(2) Know False	(3) Think True	(4) Know True	(5) All Info
Digital Literacy Video	0.001	0.002	-0.004	-0.005	-0.003
	(0.011)	(0.011)	(0.011)	(0.015)	(0.016)
Norms Intervention	-0.021	-0.027^{+}	-0.016	-0.037^{+}	-0.038^{+}
	(0.015)	(0.016)	(0.016)	(0.019)	(0.023)
Empathy Intervention	-0.001	0.013	-0.004	0.000	-0.005
	(0.016)	(0.016)	(0.016)	(0.021)	(0.023)
Popularity Intervention	-0.039**	-0.032*	-0.016	-0.016	-0.054*
	(0.014)	(0.014)	(0.015)	(0.020)	(0.021)
Northerner	0.040**	0.034*	-0.008	-0.012	0.033
	(0.015)	(0.016)	(0.016)	(0.020)	(0.023)
Yopougon	0.008	0.007	0.033**	0.035*	0.041*
	(0.012)	(0.013)	(0.013)	(0.017)	(0.018)
Port-Bouet	0.026^{+}	0.016	0.002	-0.032	0.029
	(0.015)	(0.016)	(0.017)	(0.020)	(0.024)
Ingroup Enumerator	0.028^{+}	0.028^{+}	-0.027^{+}	-0.024	0.002
	(0.015)	(0.015)	(0.015)	(0.019)	(0.022)
Constant	0.151***	0.148***	0.241***	0.278***	0.391***
	(0.017)	(0.018)	(0.019)	(0.024)	(0.028)
Observations	1500	1500	1500	1498	1500

OLS model with robust standard errors, blocking variables included.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 9: Average Treatment Effects on Affective Polarization

	(1) Full Sample	(2) Below-median Baseline Polarization	(3) Above-median Baseline Polarization
Digital Literacy Video	0.023	-0.003	0.049
	(0.032)	(0.040)	(0.052)
Norms Intervention	0.005	-0.099^{+}	0.122
	(0.046)	(0.051)	(0.078)
Empathy Intervention	0.029	-0.030	0.093
	(0.047)	(0.060)	(0.074)
Popularity Intervention	0.004	-0.071	0.092
	(0.040)	(0.050)	(0.062)
Northerner	0.055	-0.000	0.093
	(0.043)	(0.051)	(0.067)
Yopougon	0.053	0.068	0.046
	(0.035)	(0.043)	(0.056)
Port-Bouet	0.079	0.116^{+}	0.050
	(0.052)	(0.061)	(0.088)
Ingroup Enumerator	0.066	-0.003	0.128*
	(0.041)	(0.049)	(0.064)
Baseline Polarization Index	0.239***	0.024	0.226***
	(0.036)	(0.111)	(0.065)
Constant	-0.117*	-0.127	-0.202*
	(0.051)	(0.080)	(0.084)
Observations	1488	767	721

OLS model with robust standard errors, blocking variables included.

$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 11: Average Treatment Effects by Difficulty of the Information Task

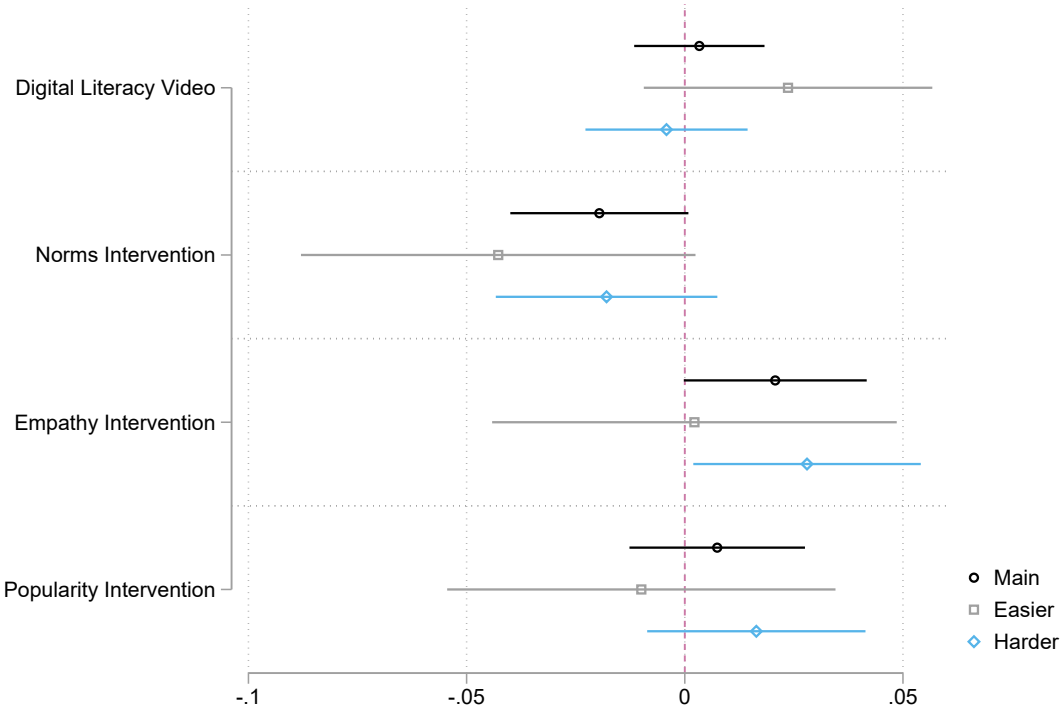


Figure 12: Baseline Levels of Polarization by Commune and Respondent Identity

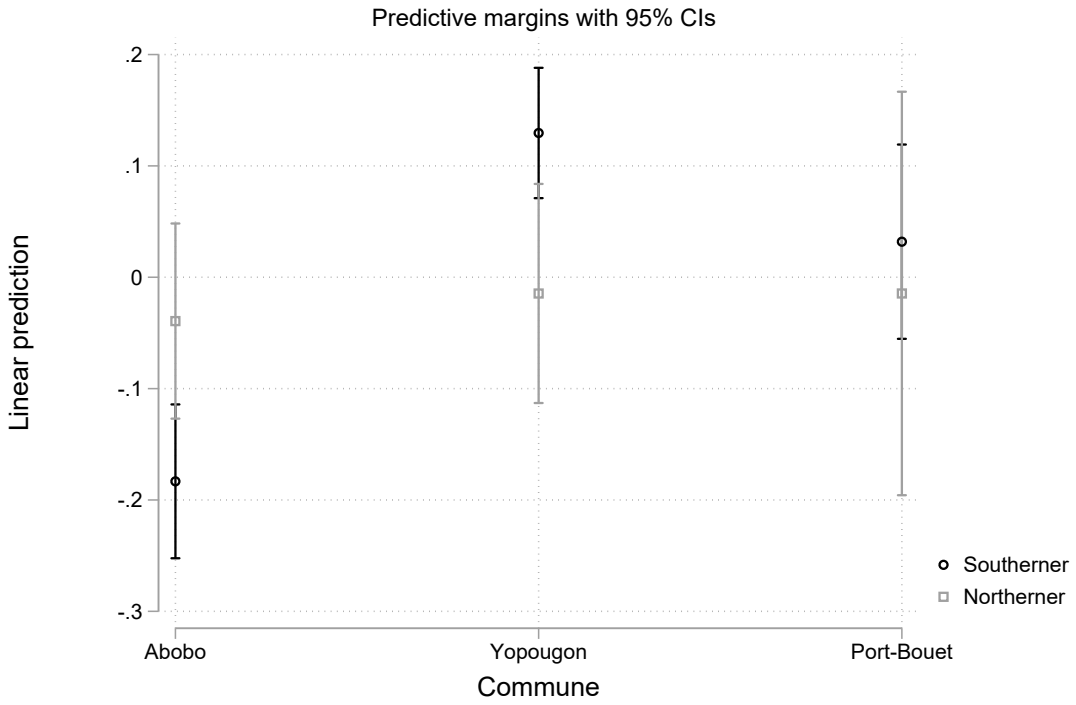
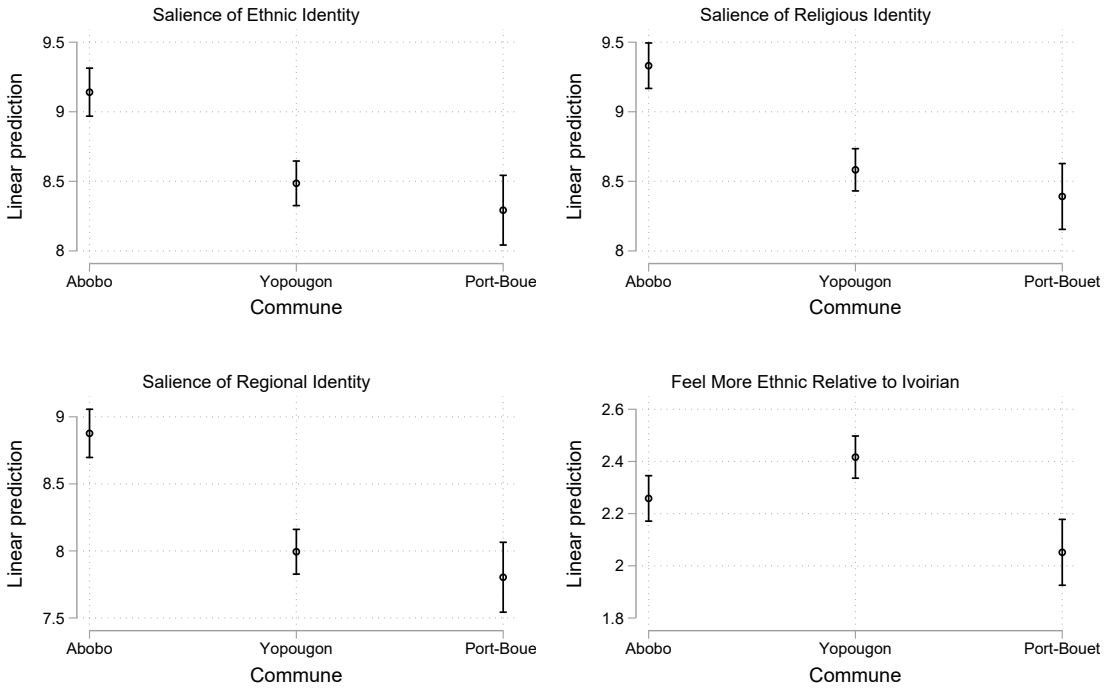


Figure 13: Measures of Identity Salience by Commune



E Ethical Considerations

The study in this paper was reviewed and approved by the University of San Diego IRB, Protocol 800894 on November 3, 2021 with a Reliance Agreement from the University of Houston. Additionally, we discussed the possibility of applying for local research ethics permission with our survey firm and among the coauthors. As of today, there is no functional ethics board for non-medical studies in Côte d'Ivoire. Additionally, the Conseil National de la Statistique (CNStat), which provides visas for statistical studies, is not yet functional either.

We consulted with our local survey firm to determine an appropriate amount for compensation. With their advice, we agreed to compensate participants with mobile transfers of 1,500 CFA (approximately 3 USD) after the first wave and 2,000 CFA (approximately 4 USD) after the second wave. It is estimated that the current minimum wage in Côte d'Ivoire is approximately 36,000 CFA a month for a 40 hour work week.[24] This is equivalent to 225 CFA per hour. Our surveys were approximately 25 minutes (wave 1) and 15 minutes (wave 2), for a total of 40 minutes. Therefore, we remunerated participants 3,500 CFA for 40 minutes of their time. This is significantly higher than the official minimum wage, and our local survey partner, FieldPro, estimated that this was a fair amount given the time and cognitive work involved.

Our participant pool was deliberately diverse on the following dimensions: ethno-regional identity, gender identity. It was restricted to 18-30 year olds with access to the internet, which was our target population. Because we sought ethno-regional diversity, our sample comprised groups that have been historically and continue to be polarized; however, polarization rather than marginalization would characterize the group dynamics within our sample. Because our study addressed intergroup relations, we relied on local partners with years of experience doing survey work in the country. The training of enumerators included sensitivity training to address potentially contentious topics such as *Ivoirité*, and the delivery method for the survey (via tablet

[24] See `https://www.ilo.org/dyn/travail/travmain.sectionReport1?p_lang=en&p_structure=1&p_year=2011&p_start=1&p_increment=10&p_sc_id=1&p_countries=CI&p_print=Y`.

with headphones in wave 1 and via email, SMS or WhatsApp in wave 2) allowed for respondent privacy.