Propaganda as Protest Prevention: How Regime Labeling Deters Citizens from Protesting—Without Persuading Them

Daniel Arnon[1], Pearce Edwards[2], and Handi Li[3]

October 14, 2020

Abstract

How do authoritarian regimes prevent protests? One strategy, which frequently accompanies the use of repression, is *labeling* regime opponents negatively in an attempt to discredit them. This paper considers two frameworks through which negative regime labels about protesters could affect citizens: through persuading them of protesters' illegitimacy, or through signaling the regime's disapproval of protest. We adjudicate the two frameworks with a survey experiment in China which varies regime responses to environmental protest. The results are consistent with the signaling disapproval framework: negative labels do not affect respondents' perceptions of protests but do affect their willingness to protest. Furthermore, these effects depend on respondents' support for the government, and suggest a polarization effect of negative labels. The findings connect research on authoritarian repression and propaganda, suggesting complementarities between the two strategies for regimes.

[1]PhD Candidate, Department of Political Science, Emory University
[2]PhD Candidate, Department of Political Science, Emory University
[3]PhD Candidate, Department of Political Science, Emory University

The authors thank Rebecca Cordell, Jennifer Gandhi, Christopher Gelpi, Adam Glynn, Joshua Kertzer, and participants in the 2020 Chinese Politics Working Group, American Political Science Association 2020 conference, and the Emory Comparative Politics Reading Group for helpful comments and suggestions. Pre-Analysis Plan registered with Evidence in Governance and Politics (ID: 20200116AB). The study was approved by Emory University's Institutional Review Board (IRB00115083).

Introduction

How do dictatorships prevent citizens from engaging in collective action? The bulk of scholarship on regimes' tactics toward protest suggests that repression—usually the direct use of state coercive power against citizens—raises the cost of collective action and either prevents or disrupts the emergence of organized opposition among the population (Svolik 2012, Sullivan 2016, Ritter and Conrad 2016, Greitens, Lee and Yazici 2019). However, the control and propagation of information is an overlooked tool in regimes' toolkit against protest. Activities of organized opposition are often complex events, and the regime can manipulate the information about these events that reaches the public (Baum and Zhukov 2015). While literature on propaganda has explored the effects of this kind of information manipulation on regime legitimacy, racial and ideological biases, and government performance (Adena et al. 2015, Yanagizawa-Drott 2014, Wedeen 2015, Huang 2015*b*, 2018), its effects have been little studied in the context of protests. We have yet to learn how the informational response of a government to a protest does (or does not) affect citizens' attitudes toward collective action.

In this paper, we fill this gap by studying whether and how authoritarian regimes' informational responses to protests affect citizens' opinions and willingness to participate in protest. We contend that regimes frequently employ these informational responses, and do so strategically. A regime decides on and propagates a message about protest events, particularly through the descriptive labels it assigns to protesters. By using negative labels—describing protests as impermissible and unacceptable—instead of labels which describe protesters sympathetically, the regime may (1) increase citizens' negative attitudes toward protesters and (2) decrease citizens' willingness to protest in the future.

We consider two theoretical frameworks through which negative labels could affect citizens. In the *persuasion* framework, a negative government label allows citizens to learn factual information about protests and thereby update both their perceptions of protesters and willingness to protest. In the *signaling disapproval* framework, the government label has no factual value for citizens learning about the protesters. Rather, the label commu-

nicates the regime's disapproval to citizens and decreases their willingness to protest. In each framework, these effects should occur most among regime supporters who are more inclined to learn from—due to prior agreement with the government's position—or comply and conform with—due to a desire to avoid conflict with the government—the government's negative labels.

We adjudicate the two theoretical perspectives with a survey experiment in China, an authoritarian regime with a certain amount of information openness but strong media control and frequent local-level protests. We randomly present respondents with negative government labels about an environmental protest event from 2012. We contrast two groups of labels: one negatively accuses protesters as criminals, the other one shows some sympathy for them. While we expect the effects of these labels vary by respondents' support for the government, it is difficult to measure respondents' sincere support for the government due to preference falsification in an authoritarian context. To elicit sincere support, we employ a list experiment and separate those respondents who falsify their support. We estimate about 58% of respondents sincerely support the Chinese regime, while 11% falsify support.

With the above empirical strategies, we identify the effect of negative government labels on respondents' perceptions of protesters and willingness to protest. We also estimate the heterogeneous effects of negative government labels by respondents' predicted sincere support for the government. We find no support for the persuasion framework: citizens do not update their perceptions of protesters, on average, when presented with a negative government label. Our findings support the signaling disapproval framework: when presented with a label describing protesters as criminals, citizens change their willingness to protest according to their estimated support for the regime. Sincere supporters, which compose the majority of our sample, become significantly less willing to protest. Thus, the regime's use of negative labels appears to deter its supporters from joining protests.

We address alternative explanations for the results in several ways. First, we confirm that the non-findings for changes in perceptions of protest are due to citizens' reactions to a

government source by testing the effects of the same labels provided by a neutral, scholarly source. We find citizens, in some cases, increase negative perceptions of protesters in response to the same label from a scholar. Second, we show that scholars' labeling of protesters does not have an impact on citizens' willingness to protest, since only the government can credibly threaten consequences. Finally, it is possible the labeling treatment contaminates our measure of government support we use based on the list experiment. We use data from an additional sample of respondents who saw only the list experiment rather than the main experiment and show that it returns similar results.

Our results suggest that, as a commonly observed yet rarely studied response to protests in authoritarian regimes, negative protest labels may serve a complementary function to the use of state coercive power: they manipulate information about existing protests to prevent future protest. However, we show that labels' effectiveness does not come through persuasion but through signaling disapproval. Citizens' perceptions do not change when they know the negative label is from the government. In contrast, they become reluctant to join protest if and only if they see the negative label from the government. Our evidence also suggests a signal of disapproval could backfire: citizens who do not support the regime may become *more* willing to protest after observing the label. In this way, we demonstrate a polarization effect of government labels consistent with recent research which suggests repression affects citizens conditional on their group membership (Nugent 2020) or regime attitudes (Young 2018). In regimes like China, which generally enjoy higher support, this presents less of a concern for the government when it uses negative labels.

The study also builds a much-needed bridge between research on repression and dissent and research on authoritarian propaganda. While an older qualitative literature speaks to the connections between propaganda and social control (Wedeen 1998, Kubik 1994, Havel and Wilson 1985), more recent scholarship has tended to divide into a focus on a "hard" repression involving state use of kinetic force to contain mass threats (Svolik 2012, Gohdes 2020, Sullivan 2016) and a "soft" repression involving the use of state media and information control to

shape citizen attitudes and behavior on issues unrelated to mass threat (Rozenas and Stukal 2019, Peisakhin and Rozenas 2018, Huang 2015*b*). The findings in this paper—consistent with Carter and Carter (2020)—return to an older perspective that "soft" repression may prevent protests parallel and complementary to "hard" repression.

Finally, in many cases, the interactions between regimes and citizens are conditioned on citizens' sincere support for the government, which is usually challenging to measure in authoritarian contexts. Using a list experiment to predict potential support for government, we also contribute to recent literature which seek to measure the underlying attitudes of citizens towards authoritarian regimes (Truex and Tavana 2019).

The Potential Effects of Regime Protest Labels

The Effect of Negative Labels on Citizen Attitudes

Authoritarian regimes often face protests from citizens who mobilize and express grievances (Kuran 1991, Sullivan 2016). In its official response to protest events, an authoritarian regime may *label* the protest or protesters: offering a description of protester characteristics. If a regime chooses a *negative* label, it questions the underlying motives of the protesters and asserts the protests are impermissible on the grounds of law-breaking (Baum and Zhukov 2015). This response could be justified through claiming protesters are criminals, terrorists, or thugs. For example, as the 2011 Arab Spring protests in Egypt gained traction, Egyptian government officials took to state media to label participation in the protests as "dangerous" on account of the presence of agitators stirring up resistance to the Mubarak regime (Lindsey 2012). In Argentina, during the country's Dirty War from 1976 to 1983, the military dictatorship referred to the victims of state repression—of whom there were up to 30,000—as "subversive elements," "delinquents," and "criminals" (Feitlowitz 2011). These labels were part of a regime strategy to persuade Argentines that repression was justified.

If a regime does not use a negative label it may instead indicate sympathy with the

protests, validating the underlying motive for the protest even if it falls short of making concessions or endorsing protesters' methods. While these words are not in themselves credible commitments to action, they can mollify protesters and buy time for the government until tangible accommodation can be made. In Egypt, after labeling protests as "dangerous" failed to prevent their escalation, dictator Hosni Mubarak held a press conference in which he struck an accommodative tone. He announced that he was "always attached to the suffering of the Egyptian people" and pledged to resolve the crisis.[4]

In this paper, we study the effects of regimes using negative labels—rather than sympathetic labels—to describe protest events. Indeed, the frequency with which regimes use negative labels about protesters suggests these labels have some strategic value. We argue that understanding the effects of these labels requires considering an additional actor in the relationship between regimes and protesters: citizens who are not (current) participants in the protest. Citizens, particularly if they sympathize with or even decide to join protests, are pivotal in shaping whether protesters gain concessions from regimes or fail to make progress on their demands (Chenoweth and Stephan 2011, Aytaç, Schiumerini and Stokes 2018). For this reason, regimes have an incentive to ensure citizen non-participants neither sympathize with nor choose to participate in protest events.

Negative labels could affect both citizens' sympathy with and willingness to participate in protests. First, by labeling protesters as criminals, a regime may persuade citizens to view these people on the streets more negatively. For example, studies suggest that a radio station that broadcast inflammatory messages against Tutsis in Rwanda encouraged violence targeting the ethnic group during the country's 1994 genocide (Yanagizawa-Drott 2014). Similarly, anti-Semitic propaganda during the Nazi dictatorship caused more denunciations of Jews (Adena et al. 2015). Second, negative labels may affect citizens' willingness to protest. Carter and Carter (2020) show that propaganda by state-owned newspapers is associated with decreased risks of protest in 31 autocracies. Huang (2018) finds that respondents in

[4]https://www.c-span.org/video/?297759-1/president-mubarak-address

China show less willingness to protest when they are shown propaganda messages intended to signal the regime's strength.

How Negative Labels Affect Citizen Attitudes

We consider two frameworks through which negative regime labels could have an effect on citizen attitudes and behavior with respect to protesters: *persuasion* and *signaling disapproval*. These frameworks draw on insights from the authoritarian propaganda literature, treating a regime's use of negative labels as a means of transmitting the regime's "social and political values" (Huang 2018). However, the frameworks diverge in their expectations for how such transmissions affect citizen attitudes. The persuasion framework suggests citizens learn from the content of negative regime labels and thereby align their beliefs and actions. The signaling disapproval framework suggests regime labels induce outward conformity and therefore a disjuncture between citizens' beliefs and their actions.

In the persuasion framework, regimes' negative labels communicate what citizens perceive as factual information about protests. Through selectively presenting information—such as labeling protesters as criminals and downplaying the alternative that protests are permissible—the regime creates a narrative about ongoing events designed to denigrate opposition while building support for the regime (Rozenas and Stukal 2019). Citizens, if they lack access to alternative sources of information about protesters against which to compare the regime's negative labels, are likely to believe the information the regime provides to them is accurate. Other research supports the notion that regime-manipulated information can shift citizens' perceptions or behavior (Rosenfeld 2018, Gandhi and Li 2019).

Thus, persuasion suggests negative labels change citizens' beliefs about protesters: increasing the perception that protesters are violent and deserving of repression. This could occur because labeling protesters as criminals identifies them as a social out-group, and citizens believe social out-groups are more deserving of state punishment (Edwards and Arnon 2019, Maoz and McCauley 2008). Furthermore, criminality could suggest protesters pose a

physical threat, leading citizens to become more supportive of the use of coercion (Hou and Quek 2019). Criminality could also suggest disruption and impact on citizens' livelihoods, which cause citizens to sour on protests (Ketchley and El-Rayyes 2019).

Yet persuasion does not affect citizens uniformly. In the context of authoritarian politics, a regime's messages tend to persuade its supporters while failing to persuade or even having the opposite effect among regime non-supporters (Peisakhin and Rozenas 2018, Gehlbach and Sonin 2014). This effect could be the result of motivated reasoning—in which supporters accept the information the regime provides because they wish it to be true—or Bayesian updating in which supporters who must trust the regime are likely to update more strongly in the direction of the message about protest. These effects likewise apply to negative regime labels about protest: citizens with prior support for the regime are more likely to update their beliefs about protesters than citizens without prior support for the regime.

Furthermore, the persuasion hypothesis expects negative regime labels affect more than citizens' attitudes about protests. Once citizens adopt negative attitudes toward protesters, they should also become less willing to protest: actions tend to follow attitudes on this issue.[5] Research on protest participation suggests a connection between individuals' sympathy for protests and willingness to participate on one hand (Pearlman 2018), and between a lack of sympathy for protests and a subsequent lack of desire to take supportive actions on the other (Zeitzoff 2018). This discussion leads to the following hypotheses:

Persuasion Hypotheses: *When a regime presents citizens with a negative label, compared to a sympathetic label, about protesters:*

- *Citizens' negative perceptions about protesters increase.*
- *Citizens' willingness to join protests decreases.*
- *These effects are larger among citizens with prior support for the regime.*

In the signaling disapproval framework, regimes' negative labels communicate not factual

[5]While there is a larger theoretical debate about the relationship between attitudes and actions, we focus here on the narrow case of attitudes and actions around protest.

information but the regime's stance toward the protests (Huang 2015*b*, Wedeen 1998). In the literature on propaganda, such regime transmissions deter citizens' willingness to oppose the regime because citizens infer that only a state with resources and capacity would convey a message so forcefully and ubiquitously. Just as importantly, citizens conclude from witnessing a negative regime message that other citizens have also witnessed the same message (Huang 2018). Thus, citizens know that undertaking the behavior the government has labeled negatively would draw the regime's criticism as well as social sanctions. As a result, the communication of regime disapproval deters protest. Disapproval of protest should be an especially credible message in stable authoritarian regimes such as China with a well-established record of preventing dissent and marshaling public support for pro-regime positions (Ritter and Conrad 2016, Greitens, Lee and Yazici 2019, Weiss 2014).

While negative labels credibly signal regime disapproval and deter protest, this framework suggests negative labels do not affect perceptions about protesters. Kubik (1994) observes from Communist Poland that such propaganda "is rigid and whatever it sells it does not sell very well", that its persuasive effects perform "very poorly under more stable circumstances" and that it has "low information value, a product of its arbitrariness and ritualism" (47). In this telling, propaganda is empty words. It does not turn citizens' hearts and minds against groups which meet the regime's disapproval, such as protesters. Rather, its effect—and likely its intent—is to induce conformity and behavioral compliance with the regime (Havel and Wilson 1985). A disjuncture between citizen attitudes and behavior results: protest is deterred, but underlying negative perceptions of protesters remain unchanged.

Similar to the persuasion framework, signaling disapproval does not affect all citizens uniformly. Some citizens may be more sensitive than others to a negative label. In particular, citizens with a strong desire to avoid conflict with the regime will have the largest decrease in willingness to protest. Participation in a public protest increases citizens' "psychological costs"—inversely to the falsification costs of Kuran (1991)—if they are reluctant to appear in disagreement with the government (Ulbig and Funk 1999). Instead, such reluctant citizens

would prefer to reduce these psychological costs through behavioral conformity (Levitan and Verhulst 2016).[6] This psychological cost could derive from fear of government criticism, fear of ostracism from peers, or an intrinsic benefit from respecting authority. We expect these psychological costs are highest among government supporters, meaning that signals of disapproval most affect supporters' willingness to protest. The following hypotheses result:

Signaling Disapproval Hypotheses: *When a regime presents citizens with a negative label, compared to a sympathetic label, about protesters:*

- *Citizens' negative perceptions about protesters do not increase.*
- *Citizens' willingness to join protests decreases.*
- *These effects are larger among citizens with prior support for the regime.*

Table 1 compares the empirical predictions of persuasion and signaling disapproval. Note the predictions differ only in their effects on citizens' negative perceptions of protesters. Therefore, negative perceptions become a discriminating test between the frameworks. If negative labels decrease willingness to join protests and do not change negative perceptions, then this would provide support for signaling disapproval. If negative labels decrease willingness to join protests and increase negative perceptions, then this would provide support for persuasion.

	Negative Perceptions	Willingness to Protest
Persuasion	+ (Supporters: +)	− (Supporters: −)
Signaling Disapproval		− (Supporters: −)

Table 1: Hypothesized Effects of Negative Labels

[6]These dynamics of conflict avoidance are even more pronounced in societies which emphasize harmony as a cultural value (Leung, Koch and Lu 2002).

Research Design

Background

We test the hypotheses with a survey experiment in China for several reasons. First, our theory relates to how governments strategically respond to protest events. To make the survey plausible for respondents, we select an authoritarian regime in which negative labels and sympathetic labels are each used with some frequency. In recent years, the number of local protests in China have increased substantially with wide variation in governmental responses ranging from expressing sympathy and accommodating demands to systematic repression.[7] For example, students, local residents and environmentalists in a city of Sichuan Province won a victory after massive protests, both violent and non-violent, against a copper smelting complex. City government officials and police blocked and talked to the marching protesters but did not label them negatively in an official statement. The US $1.6 billion project was permanently canceled in response to the protest campaign (Bradsher 2012). In contrast, a protest against a chemical plant in Zhejiang Province suffered violent repression by police. The local official media also labeled the protesters as "ill-intentioned", "illegal", and "inciting riots" (Tam 2012).[8]

Moreover, China has a vibrant, yet heavily monitored and regulated, information and media environment.[9] This means that, on one hand, information about some local protests can receive coverage in local newspapers and in social media, but on the other hand, the the government manipulates the information citizens observe about the protest. We design our survey around a salient topic within domestic politics which presents a typical case for studying government responses and information politics: environmental issues. Although protests around environmental issues do not compose a large share of protest events in China, they are usually salient and attract domestic media attention (Goebel 2019).

[7] https://www.economist.com/china/2018/10/04/why-protests-are-so-common-in-china
[8] http://china.caixin.com/2012-10-24/100451319.html
[9] The World Press Freedom Index ranks China consistently in the bottom quartile on media freedom.

Finally, conducting experiments within an authoritarian context requires addressing ethical concerns. We intentionally select environmental issues because they are salient and impact citizens' lives, but are not considered politically threatening to the central government. The environment is also an issue on which the government sometimes concedes to pressure. Indeed, the Chinese government usually tolerates public criticism of environmental problems. Thus, our survey respondents are more likely to respond without fear of retaliation. As a result, we can reasonably expect the responses we collect invite no possible repercussions for participants (Desposato 2014). We also inform respondents at the end of the survey that the researchers edited key details about environmental protests away from the actual event to ensure no enduring impact of the survey on respondents.

A Survey Experiment in China

We fielded our experiment with a survey firm in China to a sample of 3,014 respondents from February 7 to March 3, 2020.[10] Respondents were required to be adult citizens of the Chinese mainland (at least 18 out of 31 provinces) recruited through the Internet (mobile device or computer) without stratifying on any particular demographic group.[11] The sample is generally representative of the online adult population in China in terms of gender, urban residence, income, and important types of occupation. It contains a greater share of citizens with above high-school education, although this is partly because we limit our study to adults[12]. The survey experiment was designed through Qualtrics, and subjects were recruited through a Chinese online survey company which runs a crowdsourcing survey platform like Amazon Mechanical Turk.

In the survey, respondents first answer a series of pretreatment questions regarding demographic features, media consumption habits, awareness of environmental issue as well as

[10]In the discussion section, we consider the implications and external validity issue of this survey being fielded in the midst of China's coronavirus outbreak and following the Hong Kong protests.

[11]The whole survey, including survey questions, a consent form and a debrief is written completely in Mandarin Chinese.

[12]We compare our sample with the demographics in the 44th China Statistical Report on Internet Development (2019).

environmental protests. We also ask respondents to answer a non-sensitive multiple-choice question to check their attentions. We collect data on these potentially predictive covariates which allow us to assess the effectiveness of randomization.

The Label Treatment

Following the pretreatment questions, respondents are exposed to a short news story describing an environmental protest which occurred in Qidong, Jiangsu in July, 2012.[13] Respondents are also exposed to an actual picture of the event. We display the picture with the event description in order to anchor respondents with some credible observable events: although this could cause a weaker effect of the treatment, we do so because we do not intend to manipulate observable events but simply how the regime uses a label to describe those events. Following reading about the event, respondents then see a fictionalized statement which includes a government label of the event.

We randomly assign respondents into reading either a negative label or a sympathetic label from the government. Thus the treatment of interest is whether a statement describes protesters negatively or in sympathetic terms. We focus particularly on a negative label which refers to protesters as "criminals" and therefore indicates their actions are legally impermissible. Respondents are told that criminals orchestrated the protest and that their demands should not be appeased. In the sympathetic condition, respondents are informed that irrational people were the perpetrators of the protest and that their demands should be given attention. The persuasion hypotheses predicts that, compared to the sympathetic label, the negative label increases negative perceptions of protesters and decreases willingness to join protests, while the signaling disapproval hypotheses predict this label only decreases willingness to join protest.

[13] https://www.dw.com/en/chinese-city-scraps-project-after-pollution-protest/a-16129024-0

Outcome Measures

Following exposure to the environmental protest vignette with randomly-varied treatment, respondents answer a series of questions related to the outcome measures of interest: perceptions of protesters and willingness to protest. For perceptions, the first variable of interest is respondents' *attitudes toward the protesters*, asking their opinion about whether protesters were too violent and whether their actions are improper. For protest, we probe respondents' *willingness to join* similar protests, asking respondents if they would join a future protest similar to the one described in the vignette. For all the outcome questions, respondents are asked to rate on a 1 to 5 "strongly disagree" to "strongly agree" ordinal scale.

Other Variables

As referenced above, in addition to the main treatment manipulations and outcome measures we ask respondents for demographic, socioeconomic and opinion measures. These include: age, province of residence, rural locations, marital status, education level, occupation, income, party membership, news consumption, and several indicators of opinion on political and local issues such as: importance of environmental issues and protests.

Estimation of Treatment Effects

The first quantity of interest we estimate is the average treatment effect (ATE) for the negative label. This is estimated by taking the difference in means of the different outcome measures between the treatment group receiving the negative label and the treatment group receiving the sympathetic label. We then re-estimate these quantities using linear regression and include additional individual-level covariates and province fixed effects. The regression estimation for the ATEs takes the following form:

$$\text{Negative Perceptions}_i = \alpha_1 + \beta_1 \text{Label}_i + \zeta_1 X_i + \mu_p + \epsilon_i$$

$$\text{Protest Willingness}_i = \alpha_2 + \beta_2 \text{Label}_i + \zeta_2 X_i + \mu_p + \epsilon_i$$

where $Label_i$ is assignment status for the *Criminal* label for respondent i, X_i is a vector of pre-treatment covariates, μ_p is a province fixed effect and ϵ_i is a robust error term. The coefficients of interest are β_1 and β_2, the estimated effect of the negative label on negative perceptions of protesters and willingness to protest, respectively. The persuasion hypotheses expect $\beta_1 > 0$ and $\beta_2 < 0$ while the signaling disapproval hypotheses expect only $\beta_2 < 0$.

The second quantity of interest is the conditional average treatment effect (CATE) for negative labels by government support. This is estimated by interacting a measure of government support with the negative label treatment. These are included in regression estimates with individual-level covariates and province fixed effects. The regression estimation for the CATEs takes the following form:

$$\text{Negative Perceptions}_i = \alpha_1 + \beta_1 \text{Label}_i + \gamma_1 \text{Support}_i + \delta_1 \text{Label}_i \cdot \text{Support}_i + \zeta_1 X_i + \mu_p + \epsilon_i$$

$$\text{Protest Willingness}_i = \alpha_2 + \beta_2 \text{Label}_i + \gamma_2 \text{Support}_i + \delta_2 \text{Label}_i \cdot \text{Support}_i + \zeta_2 X_i + \mu_p + \epsilon_i$$

where $Label_i$ is assignment status for the *Criminal* label for respondent i and $Support_i$ is the measure of respondent support for the government. The persuasion hypotheses expect $\delta_1 > 0$ and $\delta_2 < 0$ while the signaling disapproval hypotheses expect only $\delta_2 < 0$.

The List Experiment: Measuring Government Support

As mentioned above, it is challenging to learn citizens' sincere support for the government in an authoritarian context. When measuring government support, as needed for the estimation of the CATEs, citizens may conceal their disapproval of the government due to fear of consequences (Kuran 1991). In particular, concealment would occur if the government issued negative labels about protesters: authoritarian regimes can deter expressions of opposition to government policies (Huang 2015*a*, 2018). Citizens afraid of the regime's disapproval will

then express a veneer of public support for the regime and conceal contradictory opinions (Truex and Tavana 2019). If such preference falsification exists among our respondents, it complicates our ability to interpret the effects of a regime source: we cannot be sure any CATEs we estimate reflect respondents' sincere or falsified support for the regime.

Therefore, it is necessary to estimate respondents' sincere support for the government, and thereby identify respondents who falsify their preferences. To address this problem, we conduct a list experiment with both direct and indirect questions on the sensitive item of government support (Blair and Imai 2012, Glynn 2013, Imai, Park and Greene 2015). The list experiment allows respondents to answer questions about a sensitive item in a truthful manner, embedding item among less sensitive ones.[14]

In our list experiment, we present the treatment group with a set of institutions including the local government—local governing agencies of the Chinese Communist Party—and asked how many of the listed institutions respondents find trustworthy. The control group reads a list which lacks the local government, but is otherwise identical, and is then asked how many of the institutions they find trustworthy. Specifically, we ask the treatment group:

> "Of the following institutions, how many do you think are usually reliable in protecting people's rights?
>
> - Big state-owned enterprises (SOEs)
> - Local governments
> - International human rights organizations
> - Foreign-based multinational corporations (MNCs)"

In setting the non-sensitive items, we follow recommendations by Glynn (2013) and (Blair and Imai 2012) to include two items that are likely to be negatively correlated (SOEs and international human rights organizations) and one item that is unlikely to be selected (Foreign-based MNCs) to reduce potential ceiling and floor effects. The difference in the

[14]Empirical validation studies in electoral contexts show that indirect questioning techniques in the list experiment yield estimates closer to reality (Rosenfeld, Imai and Shapiro 2016).

mean number of institutions trusted between the two groups indicates respondents' private support for the government.[15] The results, presented in Table 2, suggest sincere support in our sample is 57.9% with a 95% confidence interval of [52.1, 63.8].

The distribution of respondents' predicted probability of support is shown in Figure 1. In this paper, we use the support predicted by all our demographic and pre-treatment variables with the linear model for analysis.[16] Using the prediction from maximum likelihood estimation does not change the results. We also test for excluding different sets of predictors, and the results are similar.

	sincere support	
treatment	0.58^{***}	0.57^{***}
	(0.03)	(0.03)
expert		0.06
		(0.06)
FE & control	N	Y
Num. obs.	3014	2972

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

Table 2: Sincere Support Result from List Experiment

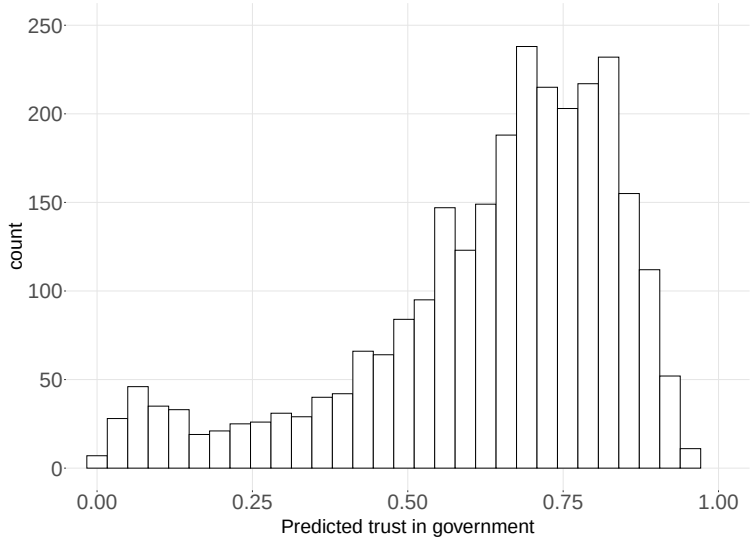
Other Treatments

At the same time, we include two additional treatment arms which assist us in better testing the mechanisms of interest. While the main analysis examines only those respondents who received the criminal label and a sympathetic label, each with a government source, there are other comparisons of interest.

First, we randomly assign some respondents to view one of the protest labels from a scholar, portrayed as a neutral source, rather than from the government. Other respondents view an identical protest label from the government, the source of interest for our main hy-

[15] In the analysis, we validate the identification of this quantity by checking the assumptions specified in Blair and Imai (2012), Imai, Park and Greene (2015).

[16] We use the R package "list" by Blair and Imai (2010). When using the linear model estimation, we drop two observations that generate predicted support greater than 1. They do not drive our results.



Notes: This figure shows the predicted support level for government using the following pre-treatment variables: gender, age, urban residence, Jiangsu, marital status, education, student, unemployment, worker, work government or public sectors, income, CCP membership, consumption of political news, social media usage, and awareness of environmental protests. It shows predictions with ML estimation. The linear model estimation gives a similar but more concentrated distribution.

Figure 1: Distribution of Predicted Support for Government

potheses. We include a scholar source for two reasons. First, it allows us to check if negative labels have effects independently of the government source, separating source effects from label content effects. Second, it allows for comparisons of the effects of negative labels which are not "politicized" with those that are. For example, the hypotheses predict heterogeneous effects by government support because the government source activates these responses. If we do not detect these heterogeneous effects with a neutral source, this builds support for the mechanism in the argument.

We also include a weaker negative label: the *number* of the protesters. This dimension captures the social identity and out-group status of the participants. In the treatment condition for this treatment arm, the protesters are described as "a very small group," suggesting to the respondents how little support the protesters elicited from the public. In the control condition, the protesters are described as "a crowd," meaning they have

significant numbers of participants and likely represent a larger group. After presenting the main results, we compare the effects of this label to the primary, criminal label of interest.

Thus, including the criminal label from the main analysis, our entire randomization scheme is a 2^3 factorial design (Government source × Criminal label × Small number label) with simple random assignment.

An example treatment vignette is shown in Figure 2, alongside the picture of the event to which respondents are exposed. All respondents are shown the introductory paragraph and the picture. Treatment manipulations are in the second paragraph.

The picture below shows a mass incident against an industrial waste pipeline project in Qidong, Jiangsu, July, 2012. Demonstrators, who worried that the project could cause water pollution, marched on the street and occupied the government building.

[The government/some scholars] stated that [**the environmental issue that people are concerned about should be given high attention/the violence committed by criminals should not be appeased**], and in this incident, [a crowd of/a very small number of] [**people irrationally/criminals violently**] stormed government and attacked the officials, which damaged social stability.

Notes: The bolded texts are the variation of the criminal label treatment. The picture is from a news report of Qidong Protest by *Deutsche Welle*.

Figure 2: Example Survey Experiment Vignette

Balance

We check the balance of the treatment and control groups across potentially predictive covariates for which we have collected data using pre-treatment questions. Balance across these covariates for the treatment arms suggests treatment assignment is independent of

potential outcomes. We present full results from balance tests in the appendix. These balance tests evaluate the difference in mean values of individual-level covariates across the different treatments. We conduct three such tests—one each for the criminal, small group, and government source treatment—and make a total of 66 comparisons.

Of these 66 comparisons, five returned significant differences between treatment groups at the $p = 0.05$ threshold: news consumption frequency for the criminal label treatment, and education level, environmental news consumption frequency, belief in environmental issues as a problem, and awareness of protests for the government source treatment. There was no imbalance for any covariates for the small group treatment. Given potential concerns about imbalance on observed covariates, especially for the government source treatment, we also report results controlling for these covariates.

Results

Distinguishing Persuasion and Signaling Disapproval

Our results show consistent support for the signaling disapproval theory but not for the persuasion theory. We find no effect of negative labels by the government on citizens' perception of protesters. Respondents do not view the protesters as more violent or more improper when they see the government label protesters as criminals, compared to the treatment conditions in which the government shows sympathy to the protesters. Columns 1-4 in Table 3 display the null results on citizens' negative perceptions of protesters. Each column alternates a baseline specification and a specification with an interaction between the criminal label treatment and respondents' predicted support for the government.

Though negative government labels do not appear to persuade citizens, we find government labels nevertheless impact citizens' willingness to participate in a protest. When people see a negative label on protesters as signaling of the government's disapproval, they process the information differently. The last column in Table 3 suggests that with a mean

	protester violent		protester improper		joining protest	
criminal	−0.05	0.04	0.04	−0.09	−0.00	0.93**
	(0.05)	(0.26)	(0.07)	(0.35)	(0.07)	(0.36)
criminal:trust		−0.15		0.21		−1.63***
		(0.43)		(0.59)		(0.61)
trust		−0.28		−0.62		−0.27
		(0.36)		(0.50)		(0.52)
FE & control	Y	Y	Y	Y	Y	Y
Num. obs.	1327	1327	1330	1330	1330	1330

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

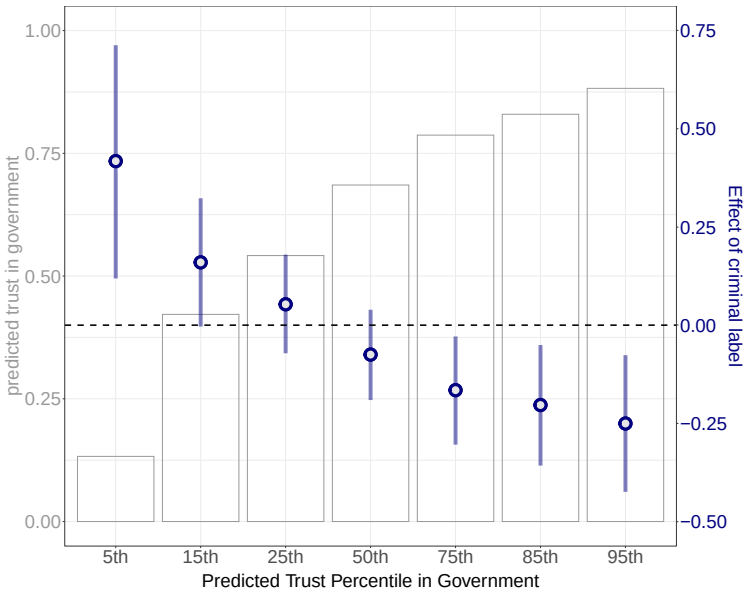
Notes: Columns 1-4 display treatment effect of criminal label (conditional on predicted trust) on perceptions on protesters in the sample of government source. Columns 5-6 display the results on willingness to join protests.

Table 3: Effect of Negative Labels (Government Source)

level of support for government (0.575 as predicted with linear model and 0.633 as predicted with ML model), the treatment effect of criminal label ranges from −0.003 to −0.058 on the willingness of joining protest.

While the effect size may seem small at the mean, Figure 3 visualizes the heterogeneity by government support depicted in the last column of Table 3. Those who trust the government to protect citizens' rights are more likely to express an unwillingness to protest. For example, at the 85th percentile of support, willingness to join protest decreases by 0.25 points (on a five-point scale). In other words, as long as citizens have a moderate level of support for government, they are less willing to participate in a protest when the government uses a negative label. As is shown in Figure 1, a majority of the respondents are predicted to have more than a medium level of support.

Figure 4 shows that when the government does not label protesters as criminals and thereby shows some sympathy to protesters, citizens are almost indifferent between whether they would join a similar protest, regardless of how much they support the government. By contrast, when the government labels protesters as criminals, regime supporters' willingness to protest is dampened, while the non-supporters express a greater willingness to go on the



Notes: The y-axis on the left shows the level of predicted support for government at 5th-95th percentile. The y-axis on the right shows estimates with 95% confidence intervals generated by simulating 10,000 random draws from the baseline interaction model at 5th-95th percentile of predicted support for government.

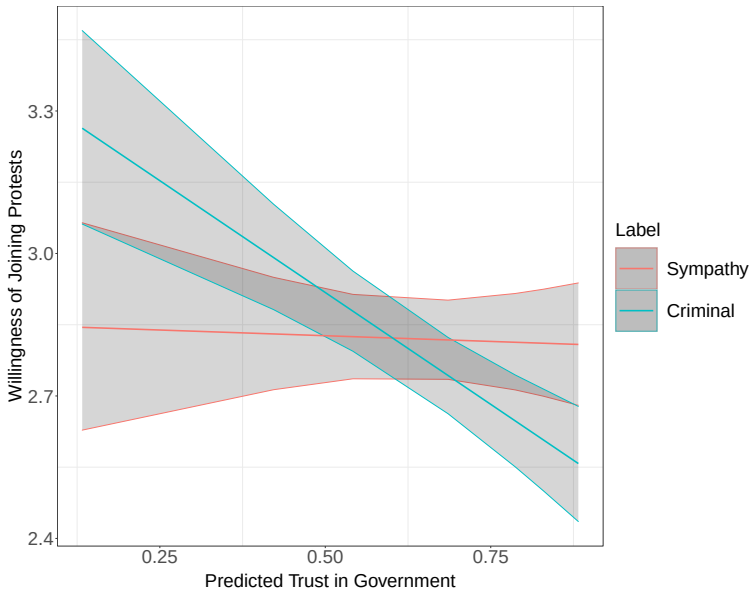
Figure 3: Effect of Criminal Label by Government Support on Joining Protests

street. These findings are consistent with signaling disapproval framework.

Testing the Government Source Mechanism

There are two main sets of alternative explanations which could account for the above results. First, it is possible that citizens' indifference to collective action events, or the respondents' lack of reaction to the treatment in our survey, explain the null results for perceptions of protest. If this is the case, we would not observe effects on negative perceptions for any protest label. Second, a different mechanism than the government source signaling disapproval, such as underlying differences in opinion about protests correlated with government support, could explain the observed results for willingness to protest. To address each of these sets of alternatives, we present results showing the effects of labels originating from a scholar source unaffiliated with the government.

We address the first alternative by comparing the effects of negative labels from a gov-



Notes: The gray bands represent 95% confidence interval. The willingness to protest is calculated by simulating 10,000 random draws from the model with criminal equals 0 and 1, respectively, conditioning on the 5th, 15th, 25th, 50th, 75th, 85th and 95th percentiles. All other variables are held at their mean values.

Figure 4: Predicted Willingness of Joining Protests by Label and Trust in Government

ernment source on perceptions of protest with the effect of negative labels by scholars on those same perceptions. Table 4 shows some negative effects for respondents' perceptions of protesters when scholars label protesters as criminals and when not conditioning on government support ($p = 0.03$ in Column 3 and $p = 0.16$ in Column 1). Because the only difference between the protest vignettes respondents read are the name of the source, we can conclude the content of the labels can affect citizens' perceptions. This insight builds support for the signaling disapproval framework: as soon as the government associates itself with a protest response, citizens no longer find the response persuasive. Furthermore, none of the interaction terms for the negative label and government source approach significance, suggesting supporters do not react differently than non-supporters to a scholar source.

We address the second alternative by comparing the effects of negative labels across sources for the willingness to protest outcome. If the signaling disapproval framework is

correct, the effects of the negative label on willingness to protest depends on the government source. That is, the government's condemnation of the protest communicates its disapproval of the protest. On the other hand, a neutral scholar's condemnation of the protest contains no implication of state disapproval. Therefore, an identical label from scholars should not affect government supporters' willingness to protest. Columns 5-6 in Table 4 provides evidence that this is the case. We also do not find any significant result excluding the interaction with predicted support for the government.

	protester violent		protester improper		joining protest	
criminal	0.07	−0.32	0.15**	−0.00	−0.05	−0.01
	(0.05)	(0.25)	(0.07)	(0.37)	(0.07)	(0.39)
criminal:trust		0.69		0.26		−0.08
		(0.44)		(0.63)		(0.68)
trust		−0.28		−0.99**		−0.92*
		(0.34)		(0.49)		(0.53)
FE & control	Y	Y	Y	Y	Y	Y
Num. obs.	1293	1293	1296	1296	1296	1296

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

Notes: Columns 1-4 display treatment effect of criminal label (conditional on predicted trust) on perceptions on protesters in the sample of scholar source. Columns 5-6 display the results on willingness to join protests.

Table 4: Effect of Negative Labels (Scholar Source)

In sum, these results from the scholar source affirm the mechanism in the argument and provide evidence against each set of alternative explanations. First, the significant main effects for labeling showing citizens increase negative perceptions of protest with a scholar source increase our confidence that the content of the labels themselves do have effects. Second, the absence of an effect of the scholar's criminal label on joining protests further validates that the mechanism of negative labeling effect must be linked with a government signal. The lack of heterogeneous effects by government support with the scholar source suggests that it is the government connection with the response to the protest that activates this polarization among respondents.

In Table 11 in the appendix, we also show the treatment effects of the negative label on the three main outcomes regardless of its source. Labeling protesters as criminals generally increases the belief that protesters are improper and an overall negative effect on willingness to protest. Yet, as we have shown above, the effect on the perceptions is driven by the scholar source while that on intention to protest is driven by the government source.

Robustness Tests

Labeling Experiment

We conduct several tests to determine the robustness of our results for the labeling experiment. First, there is a threat of potential bias from attrition. Attrition would introduce bias into our estimates if missingness in responses to outcome measures is correlated with potential outcomes. For example, respondents who support the government but also are sympathetic to protests could be reluctant to answer outcome questions when assigned to the criminal or small group treatments. We address this concern by conducting a procedure recommended in Gerber and Green (2012): regressing missingness of treatment assignment on pretreatment covariates and then comparing the F-statistic of this regression with a null distribution of F-statistics with 1,000 random treatment assignments. We conclude that missingness is independent of treatment assignment.

Because we use pre-treatment variables that are not required to answer to predict support for government, there are also missingness in the analysis with predicted support. However, missing observations are few and independent to treatment assignment, and controlling for missingness does not change our result.

Next, we determine if treatment assignment induced differences in respondents' attentiveness to the survey. Before showing the vignette, we add an attention check question in which respondents are asked to choose exactly two options for answering a non-sensitive

question.[17] Answers with exactly two options are coded as attentive, and inattentive otherwise. We find that attentiveness is balanced across treatments. And using it as a control variable or to subset the sample does not change the result.

There are also potential concerns about types of respondents who could either anticipate the purpose of the survey and give responses the researchers expect rather than sincere responses ("experts") or take the opposite approach and answer questions essentially at random ("trolls"). To account for this possibility, we classify experts as respondents who have taken five or more online surveys in the past month and classify trolls as respondents who completed the survey in less than two and a half minutes. We find no evidence these respondents are imbalanced across treatment groups, but nonetheless include covariates for experts and trolls in our analysis to account for their presence.

List Experiment Diagnostics

We begin by evaluating whether the list experiment identified sincere government supporters. That is, a list experiment to measure respondents' trust in the Chinese government is useful if there are indeed respondents in our sample who indicate a publicly-revealed preference for the government while not indicating such a preference privately. To ascertain the extent of such preference falsification—-called desirability bias in the list experiment context—we compare the difference in estimated government support from the list experiment's treatment effect with responses to a direct question about trust in local government asked only to the respondents assigned to control in the list experiment.[18]

To make this comparison, we first add a value of one to the control group's list experiment response if they indicated trust in local government on the direct question, and add no value if they did not indicate trust in local government. We call this the *publicly-revealed preference* for the government. We then take the difference between the public preference

[17] Specifically, we ask respondents to choose two out of five sources of air pollution that they believe are important.

[18] This direct question appears for the control group after they have completed the list experiment to avoid introducing post-treatment bias in the list experiment.

and the treatment group's list experiment response – the *privately-revealed preference* for the government – to estimate preference falsification:

$$\begin{aligned}\hat{\tau}_{\text{falsification}} &= \underbrace{\frac{1}{N_0}\sum_{i=1}^{N}[(1-T_i)(Z_i)]}_{\text{Public Preference}} - \underbrace{(\frac{1}{N_1}\sum_{i=1}^{N}T_iY_i - \frac{1}{N_0}\sum_{i=1}^{N}[(1-T_i)(Y_i)])}_{\text{Private Preference}} \\ &= \frac{1}{N_0}\sum_{i=1}^{N}[(1-T_i)(Y_i+Z_i)] - \frac{1}{N_1}\sum_{i=1}^{N}T_iY_i\end{aligned}$$

where N_0 is the number of respondents in the control group, N_1 is the number of respondents in the treatment group, T_i is an indicator for treatment for respondent i, Y_i is the count of institutions respondent i trusts in the list experiment, and Z_i is an indicator for respondent i's trust in local government from the direct question.

If the estimator $\hat{\tau}_{\text{falsification}}$ takes a positive and significant value, indicating the mean publicly-expressed trust in local government is greater than the mean privately-expressed trust in local government, we infer our sample contains nonzero preference falsification. If the estimator takes a value indistinguishable from zero, indicating similar trust in local government in the direct question and the list experiment, we cannot conclude our sample contains preference falsification. Detecting nonzero preference falsification would validate the use of a list experiment, as we are interested in estimating heterogeneous effects among *sincere* government supporters: those respondents who indicate support for the government publicly and privately.

Using the estimator $\hat{\tau}_{\text{falsification}}$, we estimate the share of falsifiers in the sample as 10.2% with a 95% confidence interval of [3.5, 16.9]. This latter estimate confirms nonzero preference falsification in the sample, suggesting that our use of respondents' estimated government trust from the list experiment is preferred to using respondents' direct expression of government trust. Results from this test are reported in Table 5.

Next, we conduct list experiment diagnostic tests. A first test is to determine whether

	preference falsification	
treatment	0.10^{***}	0.11^{***}
	(0.03)	(0.03)
expert		-0.08
		(0.07)
FE & control	N	Y
Num. obs.	3014	2972

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

Table 5: Preference Falsification Result from List Experiment

there are *design effects* in the experiment. This would be the case if respondents' treatment assignment affected their responses to the control items in the list (trust in state-owned enterprises, international human rights organizations, and foreign multinational corporations). We conduct a test for design effects following Blair and Imai (2012) and find no evidence of them: the adjusted p-value of the test is 1. This increases our confidence in our ability to identify the proportion of those who sincerely trust the government.

A second test is accounting for whether there are *liars* in the sample. That is, do all respondents give truthful answers to the government trust item in the list experiment? While the baseline list experiment results used in the analysis assume no liars, we can relax the assumption to account for the possibility of "ceiling effects": that respondents in the treatment group would be unwilling to indicate trust in the government when answering affirmatively to all other items in the list. The procedure, from Blair and Imai (2012), adjusts estimates of the relationship between covariates and government trust according to the estimated proportion of liars. We find no significant differences between estimates adjusting for ceiling effects and the unadjusted estimates.

Finally, we consider the possibility that we introduce post-treatment bias into our estimates by conditioning treatment effects on predicted government support. This could be the case because the list experiment in our original sample follows the labeling experiment. To address this concern, we fielded a second survey with 502 respondents and no labeling experiment. Estimated sincere trust in the government from the list experiment in this sam-

ple is 56.7%, almost identical to the original sample. Additionally, we find no evidence that any pretreatment covariate predicts government support differently across the two samples: when pooling the samples and interacting a survey indicator variable with each predictive covariate, the interaction term is never significant. Furthermore, none of the labeling treatments in the original sample predict the list experiment result. This allays concerns that the labeling experiment could affect respondents' government trust according to the values of pretreatment covariates. The results of this test in the appendix, depicted in Figure 5.

Sample Selection

Since our survey sample was recruited through an online platform and the recruiting process cannot be monitored by researchers, we need to make sure that the sample is not biased at least compared to the population who have access to Internet in China. We compare the important demographics in our sample with the Internet population data from *The 45th Statistical Report on Internet Development in China* by the China Internet Network Information Center (CNNIC) in 2020. The comparison is shown in Table in Appendix. The sample we use is likely to be biased towards female, age between 20 and 29, and higher education.[19]

We therefore use reweighting to check the oversampling for the three groups. The weights are calculated by sample proportion divided by population proportion. The reweighted results are shown in Appendix. They hold the same with results from our main models.

Transparency and Pre-Analysis Plan

Here we briefly list the similarities and differences between our findings and our original expectations for effects. In general, we anticipated that negative labels and a government source would (1) increase negative perceptions of protests, and (2) decrease willingness to

[19]Note that our sample only include adults while the official statistics of population do not make adulthood distinction. So, the actual proportion of people with lower education degree would have been higher if we include non-adults. However, it is still likely that we oversampled people with a degree of higher education.

protest. Our expectations were generally confirmed in that negative labels did have effects on perceptions and willingness to protest. However, these effects depended on the source. Negative labels increased negative perceptions with a scholar source, while decreasing willingness to protest with a government source—at least among the majority of the sample who trust the government. Conversely, we did not find that sources had effects independent of the content of the labels they gave. The results paint a more nuanced picture of the conditions under which labels affect perceptions than we anticipated.

Discussion and Conclusion

Authoritarian regimes' use of negative labels in response to protest affect public attitudes. From a survey experiment in China, we have shown negative labels affect citizens' willingness to protest, and that these effects are conditional on prior support for the government. We find no ability for a government source to affect citizens' perceptions of protesters, even as a neutral source is able to affect these perceptions. The combination of these findings supports the proposed *signaling disapproval* framework for government labels, in which negative labels do not communicate factual information to citizens but rather indicate the government's official position toward certain types of protest.

Finding only conditional effects from the government's use of negative labels points to the limitation of information-based responses to protest in authoritarian regimes. While evidence suggests that the facts citizens perceive about protests in both authoritarian regimes (Hou and Quek 2019) and democracies (Edwards and Arnon 2019) consistently increase support for state use of force, citizens appear less persuaded by the labels assigned to observed facts which are otherwise held constant. This finding suggests the need to identify the effects of labels—such as "terrorism"—on citizens' attitudes in democracies to determine if such labels are actually consequential for opinion. If respondents are more likely to label social out-groups as terrorists,[20] it is possible that support for repression of terrorism is a function

[20]See, for example, Huff and Kertzer (2018).

of the terrorist's identity more than the terrorist label itself.

We also find some evidence of a backfire effect: respondents who least support the government are more likely to protest after seeing a negative government label. However, we note that, based on Figure 1, only a small minority of citizens fall into this category. This suggests a small *polarization* effect of government messages about protest (Peisakhin and Rozenas 2018), in which the government is only partially able to deter participation in additional protests. Finding backfire and polarization effects from negative government labels could connect to research on the backfire (Aytaç, Schiumerini and Stokes 2018) and polarization (Nugent 2020) effects of repression, suggesting another link between literature on propaganda and repression which scholars should consider.

This paper has also linked literature on repression and propaganda by showing that government labels form part of what Davenport (2015) calls the "problem depletion" strategy of states against challengers such as protesters. In this strategy, "ordinary citizens and government opponents are bombarded with information indicating that they should not support challengers and that the dissidents are losing the confrontation" (Davenport 2015, 28). Yet "problem depletion" seems only to have modest effects when governments are the source of such information bombardment. The lack of government-source effects compared to neutral-source effects could be one reason why governments attempt to disseminate messages through "grey propaganda" and "black propaganda": pro-government messages which appear to originate from another source such as a think tank or even the opposition.

There could be external validity concerns since the experiment occurred during the breakout of coronavirus and close to Hong Kong protests in China. In February and early March 2020, China was still in the midst of its unprecedented policy response to the coronavirus outbreak which originated in Hubei Province. While our design has accounted for possible geographic variation in responses, and while randomization should leave our treatment effects free of potential omitted variables correlated with the outbreak, there remain questions about external validity. Do citizens in an authoritarian regime respond similarly to protests

about local issues when faced with a national crisis such as coronavirus than when no such crisis exists? While some literature has suggested that the Chinese government can harness public opinion in times of crisis—particularly related to foreign disputes (Weiss 2014)—the polarizing effects of a government source in this study suggest a lack of unity behind the Chinese Communist Party. Similarly, the escalation of the Hong Kong protests at the end of 2019 may have an impact on how people view the state-society relationship in China.

What remains unexplained is the ubiquity of governments' use of labeling. If such labels have only modest effects on public opinion when they originate from the government, then scholars should consider and test arguments related to the other purposes these labels might serve when governments employ them. In particular, government labels could be a means of communicating to other agents in a political regime. Labeling a group of protesters "criminals" or a "fringe group" could be a signal or trigger for a bureaucratic process to empower military agents and green-light repression. On the other hand, labeling a group of protesters as simply irrational or representative of society could provide a cue to local officials that protests are to be tolerated and leveraged for purposes such as information collection (Chen and Xu 2017, Lorentzen 2013). Labeling could also be a means of regime hard-liners staking out a repressive policy position in a competition with more soft-line regime insiders (Thomson 2017). Future research should consider these possible explanations.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa and Ekaterina Zhuravskaya. 2015. "Radio and the rise of the Nazis in prewar Germany." *Quarterly Journal of Economics* 130(4):1885–1939.

Aytaç, S Erdem, Luis Schiumerini and Susan Stokes. 2018. "Why do people join backlash protests? Lessons from Turkey." *Journal of Conflict Resolution* 62(6):1205–1228.

Baum, Matthew A and Yuri M Zhukov. 2015. "Filtering revolution: Reporting bias in

international newspaper coverage of the Libyan civil war." *Journal of Peace Research* 52(3):384–400.

Blair, Graeme and Kosuke Imai. 2010. "list: Statistical Methods for the Item Count Technique and List Experiment." Available at The Comprehensive R Archive Network (CRAN).
URL: *https://CRAN.R-project.org/package=list*

Blair, Graeme and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20:47–77.

Bradsher, Keith. 2012. "Bolder Protests Against Pollution Win Project's Defeat in China.".
URL: *https://www.nytimes.com/2012/07/05/world/asia/chinese-officials-cancel-plant-project-amid-protests.html*

Carter, Erin Baggott and Brett L Carter. 2020. "Propaganda and Protest in Autocracies." *Working paper* .

Chen, Jidong and Yiqing Xu. 2017. "Why Do Authoritarian Regimes Allow Citizens to Voice Opinions Publicly?" *The Journal of Politics* 79(3):792–803.

Chenoweth, Erica and Maria Stephan. 2011. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. New York: Columbia University Press.

Davenport, Christian. 2015. *How Social Movements Die: Repression and Demobilization of the Republic of New Africa*. Cambridge University Press.

Desposato, Scott. 2014. "Ethics and research in comparative politics.".
URL: *https://www.washingtonpost.com/news/monkey-cage/wp/2014/11/03/ethics-and-research-in-comparative-politics/*

Edwards, Pearce and Daniel Arnon. 2019. "Violence on Many Sides: Framing Effects on Protest and Support for Repression." *British Journal of Political Science* Forthcoming.

Feitlowitz, Marguerite. 2011. *A Lexicon of terror: Argentina and the legacies of torture*. Oxford: Oxford University Press.

Gandhi, Jennifer and Handi Li. 2019. "Data Manipulation and its Effect on Citizens' Views: Evidence from a Survey Experiment in China." *Working paper* .

Gehlbach, Scott and Konstantin Sonin. 2014. "Government Control of the Media." *Journal of Public Economics* 118:163–171.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis and Interpretation*. New York: W.W. Norton.

Glynn, Adam N. 2013. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77(S1):159–172.

Goebel, Christian. 2019. Social unrest in China: a bird's-eye view. In *Handbook of Protest and Resistance in China*. Edward Elgar Publishing.

Gohdes, Anita R. 2020. "Repression Technology: Internet Accessibility and State Violence." *American Journal of Political Science* Forthcoming.

Greitens, Sheena Chestnut, Myunghee Lee and Emir Yazici. 2019. "Counterterrorism and Preventive Repression: China's Changing Strategy in Xinjiang." *International Security* 44(3):9–47.

Havel, Vaclav and Paul Wilson. 1985. "The power of the powerless." *International Journal of Politics* 15(3/4):23–96.

Hou, Yue and Kai Quek. 2019. "Violence Exposure and Support for State Use of Force in a Non-Democracy." *Journal of Experimental Political Science* 6:120–130.

Huang, Haifeng. 2015*a*. "Propaganda as signaling." *Comparative Politics* 47(4):419–444.

Huang, Haifeng. 2015*b*. "A War of (Mis)Information: The Political Effects of Rumors and Rumor Rebuttals in an Authoritarian Country." *British Journal of Political Science* 47:283–311.

Huang, Haifeng. 2018. "The Pathology of Hard Propaganda." *Journal of Politics* 80(3):1034–1038.

Huff, Connor and Joshua D. Kertzer. 2018. "How the Public Defines Terrorism." *American Journal of Political Science* 62(1):55–71.

Imai, Kosuke, Bethany Park and Kenneth F Greene. 2015. "Using the predicted responses from list experiments as explanatory variables in regression models." *Political Analysis* 23(2):180–196.

Ketchley, Neil and Thoraya El-Rayyes. 2019. "Unpopular Protest: Mass Mobilization and Attitudes to Democracy in Post-Mubarak Egypt." *Journal of Politics* Forthcoming.

Kubik, Jan. 1994. *The Power of Symbols Against the Symbols of Power: The Rise of Solidarity and the Fall of State Socialism in Poland*. University Park: Penn State University Press.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(1):7-48.

Leung, Kwok, Pamela Tremain Koch and Lin Lu. 2002. "A dualistic model of harmony and its implications for conflict management in Asia." *Asia Pacific Journal of Management* 19(2-3):201–220.

Levitan, Lindsey C and Brad Verhulst. 2016. "Conformity in groups: The effects of others' views on expressed attitudes and attitude change." *Political Behavior* 38(2):277–315.

Lindsey, Ursula. 2012. Revolution and Counterrevolution in the Egyptian Media. In *The*

Journey to Tahrir: Revolution, Protest and Social Change in Egypt, ed. Jeannie Sowers and Chris Toensing. London: Verso Books pp. 41–46.

Lorentzen, Peter J. 2013. "Regularizing Rioting: Permitting Public Protest in an Authoritarian Regime." *Quarterly Journal of Political Science* 8:127–158.

Maoz, Ifat and Clark McCauley. 2008. "Threat, Dehumanization, and Support for Retaliatory Aggressive Policies." *Journal of Conflict Resolution* 52(1):93–116.

Nugent, Elizabeth R. 2020. "The Psychology of Repression and Polarization." *World Politics* 72(2):291–334.

Pearlman, Wendy. 2018. "Moral Identity and Protest Cascades in Syria." *British Journal of Political Science* 48(4):877–901.

Peisakhin, Leonid and Arturas Rozenas. 2018. "Electoral Effects of Biased Media: Russian Television in Ukraine." *American Journal of Political Science* 62(3):1165–1177.

Ritter, Emily Hencken and Courtenay R. Conrad. 2016. "Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression." *The American Political Science Review* 110(1):85–99.

Rosenfeld, Bryn. 2018. "The popularity costs of economic crisis under electoral authoritarianism: evidence from Russia." *American Journal of Political Science* 62(2):382–397.

Rosenfeld, Bryn, Kosuke Imai and Jacob N Shapiro. 2016. "An empirical validation study of popular survey methodologies for sensitive questions." *American Journal of Political Science* 60(3):783–802.

Rozenas, Arturas and Denis Stukal. 2019. "How autocrats manipulate economic news: Evidence from Russia's state-controlled television." *The Journal of Politics* 81(3):982–996.

Sullivan, Christopher M. 2016. "Undermining Resistance: Mobilization, Repression, and the Enforcement of the Political Order." *Journal of Conflict Resolution* 60(7):1163–1190.

Svolik, Milan. 2012. *The Politics of Authoritarian Rule*. New York: Cambridge University Press.

Tam, Fiona. 2012. "Scuffles as Ningbo residents step up chemical plant protest.".
URL: *https://www.scmp.com/news/china/article/1071428/scuffles-ningbo-residents-step-chemical-plant-protest*

Thomson, Henry. 2017. "Repression, Redistribution and the Problem of Authoritarian Control: Responses to the 17 June Uprising in Socialist East Germany." *East European Politics and Societies and Cultures* 31(1):34–42.

Truex, Rory and Daniel L. Tavana. 2019. "Implicit Attitudes Toward an Authoritarian Regime." *The Journal of Politics* 81(3):1014–1027.

Ulbig, Stacy G and Carolyn L Funk. 1999. "Conflict avoidance and political participation." *Political Behavior* 21(3):265–282.

Wedeen, Lisa. 1998. "Acting 'As If': Symbolic Politics and Social Control in Syria." *Comparative Studies in Society and History* 40(3):503–523.

Wedeen, Lisa. 2015. *Ambiguities of domination: Politics, rhetoric, and symbols in contemporary Syria*. University of Chicago Press.

Weiss, Jessica C. 2014. *Powerful Patriots: Nationalist Protest in China's Foreign Relations*. Oxford: Oxford University Press.

Yanagizawa-Drott, David. 2014. "Propaganda and conflict: Evidence from the Rwandan genocide." *The Quarterly Journal of Economics* 129(4):1947–1994.

Young, Lauren E. 2018. "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review* Forthcoming:1–16.

Zeitzoff, Thomas. 2018. "Anger, legacies of violence, and group conflict: An experiment in post-riot Acre, Israel." *Conflict Management and Peace Science* 35(4):402–423.

Appendices

Summary Statistics

Table 6: Summary of Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
female	3,012	0.531	0.499	0	1	1
age	3,014	27.156	6.306	18	26	90
student	3,013	0.212	0.409	0	0	1
rural	3,009	0.656	0.475	0	1	1
married	2,892	0.357	0.479	0	0	1
education	2,961	4.436	0.923	1	5	6
occupation	3,013	3.910	2.996	1	3	13
unemployed	3,013	0.031	0.173	0	0	1
government_work	3,012	0.230	0.421	0	0	1
worker	3,013	0.056	0.230	0	0	1
Jiangsu	3,007	0.079	0.269	0	0	1
income	3,013	3.208	1.205	1	3	8
CCP	3,012	0.176	0.381	0	0	1
sports_freq	3,010	2.784	0.781	1	3	4
news_freq	2,939	2.648	0.758	1	3	4
envr_freq	2,950	2.591	0.704	1	3	4
media_type	3,011	0.063	0.244	0	0	1
envr_problem_pre	3,010	3.865	0.631	1	4	5
econ_dev_pref	2,980	2.031	0.826	1	2	5
protest_aware	2,986	2.708	0.887	1	3	4
attcheck	3,014	0.836	0.371	0	1	1
expert	2,979	0.071	0.257	0	0	1
troll	3,014	0.003	0.058	0	0	1
violence	3,007	3.922	0.869	1	4	5
attention	3,014	0.989	0.104	0	1	1
improper	3,014	3.875	1.228	1	4	5
arrest_approp	3,014	4.046	1.091	1	4	5
cancel_approp	3,014	4.220	1.015	1	5	5
envr_problem_post	3,014	4.438	0.821	1	5	5
join	3,014	2.786	1.296	1	3	5
gov_response_no	2,973	1.865	1.233	1	1	5
gov_response_police	2,973	4.285	0.936	1	5	5
gov_response_arrest	2,968	3.981	1.115	1	4	5
gov_response_arrestall	2,966	2.384	1.322	1	2	5
best_response	2,964	2.102	0.444	1	2	4
Tgov	3,014	0.505	0.500	0	1	1
Tcriminal	3,014	0.508	0.500	0	1	1
Tsmgrp	3,014	0.490	0.500	0	0	1
list_treat	3,014	0.477	0.500	0	0	1
loc_gov_trust	1,575	0.681	0.466	0	1	1

Balance Tests

Table 7: Balance for Government Source Treatment

variable	diff_mean	t	se	p
female	0.016	0.8222	0.0195	0.4111
age	0.2967	1.2581	0.2358	0.2086
rural	0.0211	1.1466	0.0184	0.2518
Jiangsu	-0.0013	-0.1179	0.0108	0.9062
married	-0.018	-0.9589	0.0188	0.3378
education	0.0713	2.0028	0.0356	0.0454
student	0.0217	1.3868	0.0157	0.1658
unemployed	-0.0046	-0.6602	0.007	0.5093
worker	-0.0036	-0.4077	0.0089	0.6836
government_work	-0.0127	-0.7612	0.0166	0.4467
income	-0.0579	-1.2408	0.0466	0.2149
CCP	0.0117	0.7838	0.0149	0.4333
sports_freq	-0.0409	-1.3461	0.0304	0.1785
news_freq	-0.0176	-0.5974	0.0295	0.5504
envr_freq	-0.0603	-2.1982	0.0274	0.0281
media_type	0.0077	0.8341	0.0092	0.4044
envr_problem_pre	-0.0511	-2.097	0.0244	0.0362
econ_dev_pref	-0.0281	-0.8672	0.0325	0.386
protest_aware	-0.0698	-2.0273	0.0345	0.0428
troll	-0.0016	-0.8437	0.0019	0.399
expert	-0.0024	-0.2469	0.0099	0.805
attcheck	0.0076	0.537	0.0141	0.5913

Table 8: Balance for Criminal Labeling Treatment

variable	diff_mean	t	se	p
female	-0.0164	-0.8395	0.0195	0.4013
age	0.0864	0.3654	0.2365	0.7149
rural	0.0146	0.79	0.0184	0.4297
Jiangsu	0.0069	0.6352	0.0108	0.5254
married	0.0157	0.834	0.0188	0.4044
education	0.0703	1.9744	0.0356	0.0485
student	-0.0112	-0.7146	0.0157	0.475
unemployed	-0.002	-0.2824	0.007	0.7777
worker	0.0125	1.4121	0.0088	0.1582
government_work	-0.0171	-1.027	0.0166	0.3046
income	0.0794	1.7021	0.0466	0.089
CCP	-0.0149	-0.9966	0.0149	0.3191
sports_freq	-0.0157	-0.5147	0.0305	0.6069
news_freq	-0.0593	-2.006	0.0295	0.0451
envr_freq	-0.0233	-0.8502	0.0274	0.3953
media_type	0.0024	0.2598	0.0092	0.7951
envr_problem_pre	0.0129	0.5271	0.0244	0.5982
econ_dev_pref	0.0188	0.578	0.0325	0.5633
protest_aware	-0.0236	-0.6841	0.0345	0.4941
troll	0.0014	0.777	0.0019	0.4373
expert	-0.0033	-0.3319	0.0099	0.74
attcheck	0.0095	0.6707	0.0141	0.5025

Table 9: Balance for Small-Group Labeling Treatment

variable	diff_mean	t	se	p
female	-0.0016	-0.0815	0.0195	0.935
age	0.2365	1.0013	0.2362	0.3169
rural	0.0129	0.6997	0.0184	0.4843
Jiangsu	0.0057	0.5293	0.0108	0.5967
married	-0.0176	-0.9372	0.0188	0.3488
education	-0.0262	-0.7366	0.0356	0.4615
student	-0.0099	-0.6282	0.0157	0.53
unemployed	0.0091	1.3059	0.007	0.1918
worker	0.0051	0.5702	0.0089	0.5687
government_work	-0.0106	-0.6366	0.0166	0.5245
income	-0.0416	-0.8913	0.0467	0.3729
CCP	-0.0196	-1.3137	0.0149	0.1892
sports_freq	-0.003	-0.0977	0.0304	0.9222
news_freq	0.009	0.305	0.0296	0.7604
envr_freq	-0.0367	-1.3371	0.0274	0.1814
media_type	-0.0078	-0.8439	0.0092	0.3989
envr_problem_pre	0.0315	1.2911	0.0244	0.1969
econ_dev_pref	0.0494	1.5244	0.0324	0.1276
protest_aware	-0.0014	-0.0396	0.0345	0.9684
troll	1e-04	0.028	0.0019	0.9776
expert	0.0069	0.6981	0.0099	0.4852
attcheck	-0.0013	-0.0891	0.0141	0.929

Table 10: Balance for List Experiment

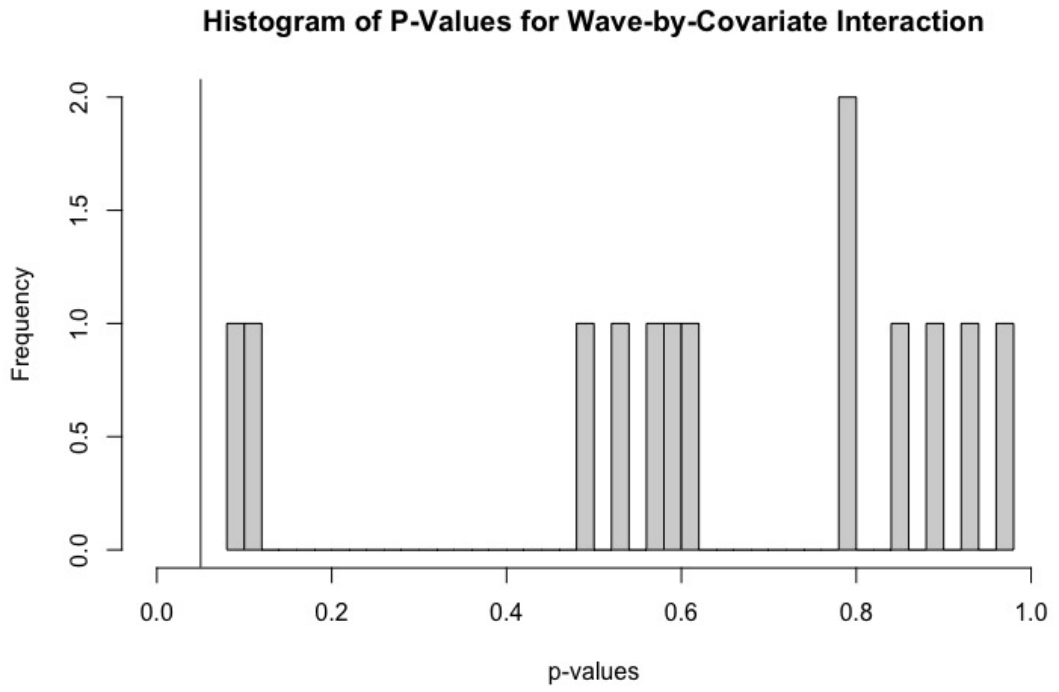
variable	diff_mean	t	se	p
female	-0.0258	-1.3217	0.0195	0.1865
age	-0.0987	-0.418	0.2361	0.676
rural	0.0026	0.1411	0.0185	0.8878
Jiangsu	0.0045	0.4125	0.0108	0.6801
married	0.0096	0.509	0.0188	0.6109
education	-0.0168	-0.4711	0.0357	0.6377
student	-0.0088	-0.5618	0.0157	0.5743
unemployed	0.0034	0.4915	0.007	0.6232
worker	0.0067	0.7534	0.0089	0.4513
government_work	0.0108	0.6499	0.0166	0.5159
income	0.0689	1.4767	0.0467	0.14
CCP	-0.0139	-0.933	0.0149	0.351
sports_freq	-0.0098	-0.3218	0.0305	0.7476
news_freq	0.0031	0.1045	0.0296	0.9168
envr_freq	0.0513	1.8708	0.0274	0.0616
media_type	-0.0105	-1.1354	0.0092	0.2564
envr_problem_pre	0.0474	1.9367	0.0245	0.053
econ_dev_pref	0.0023	0.0722	0.0325	0.9425
protest_aware	0.0458	1.3255	0.0345	0.1853
troll	0.0017	0.9035	0.0019	0.3664
expert	0.0292	2.9343	0.01	0.0034
attcheck	0.0035	0.246	0.0141	0.8057

Robustness Tests

	protester violent		protester improper		joining protest	
criminal	0.00	0.01	0.08*	0.09*	−0.09**	−0.06
	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)
social news		0.04*		0.04		0.10***
		(0.02)		(0.03)		(0.03)
education		−0.02		0.01		−0.04
		(0.02)		(0.03)		(0.03)
Num. obs.	3007	2843	3014	2849	3014	2849

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

Table 11: Effect of Negative Labels (Both Sources)



Notes: Histogram indicates frequency of p-values from tests of list experiment wave-by-covariate interaction terms for each covariate collected across the two surveys. Vertical gray line indicates threshold of significance below which we would reject the null hypothesis of no relationship between wave and predictive value of covariate for government support. No p-values fall below this threshold.

Figure 5: P-Values from List Experiment Wave-by-Covariate Interaction Terms

Results with Small-Group Label

	protester violent		protester improper		joining protest	
small group	−0.36	−0.39	−0.60*	−0.58*	0.21	0.21
	(0.25)	(0.25)	(0.34)	(0.35)	(0.36)	(0.36)
small group:trust	0.69*	0.73*	1.12*	1.07*	−0.38	−0.43
	(0.42)	(0.43)	(0.57)	(0.59)	(0.60)	(0.61)
social news		−0.02		0.07		0.08
		(0.04)		(0.05)		(0.05)
education		0.01		0.04		0.00
		(0.03)		(0.04)		(0.04)
pollution prior		0.01		0.04		0.26***
		(0.04)		(0.05)		(0.06)
prefer development		−0.01		0.01		−0.06
		(0.03)		(0.04)		(0.04)
FE & control	N	Y	N	Y	N	Y
Num. obs.	1377	1327	1380	1330	1380	1330

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

Table 12: Effect of Small-Group Labels (Government Source)

	protester violent		protester improper		joining protest	
small group	0.10	0.24	0.08	−0.08	−0.02	0.10
	(0.25)	(0.25)	(0.35)	(0.37)	(0.38)	(0.39)
small group:trust	−0.01	−0.26	−0.15	0.16	−0.02	−0.19
	(0.43)	(0.44)	(0.61)	(0.63)	(0.65)	(0.68)
social news		0.04		0.00		0.06
		(0.04)		(0.05)		(0.06)
education		−0.02		0.03		0.01
		(0.03)		(0.04)		(0.05)
pollution prior		0.05		−0.09		0.15**
		(0.04)		(0.06)		(0.06)
prefer development		0.01		−0.08*		−0.06
		(0.03)		(0.04)		(0.04)
FE & control	N	Y	N	Y	N	Y
Num. obs.	1348	1293	1351	1296	1351	1296

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

Table 13: Effect of Small-Group Labels (Scholar Source)

Sample Demographics

Demographics	Sample (Adult)	Population
Gender		
Female	53.1	48.1
Male	46.9	51.9
Residence		
Rural	34.4	28.2
Urban	65.6	71.8
Age*		
20-29	65.3	26.7
30-39	26.1	25.8
40-49	3.1	21.8
50-59	0.6	12.7
60-	0.3	8.3
Education*: not comparable as sample only includes adults		
Primary school or under	0.3	17.2
Junior high school	2.9	41.1
Senior high school/Vocational school	13.4	22.2
Junior college and above	83.3	19.5
Occupation		
Government/Public sector employee	6.7	2.8
Student	21.2	26.9
Manufacturing worker	1.4	2.6
Professional	8.2	6.0
Self-employed	12.2	22.4
Rural migrant worker	0.9	4.2
Farmer	0.1	6.3
Corporate management	11.3	2.9
Corporate office worker	27.3	8.0
Service worker	3.4	4.4
Retired	0.1	4.7
Unemployed	3.1	8.8

Notes: Percentages of the demographic features. The population data is derived from *The 45th Statistical Report on Internet Development in China* (2020) by the China Internet Network Information Center (CNNIC). The age structure of population is reweighted by adult netizens' number, which is from *Report on Non-Adult Usage of Internet in China (2019).*

Table 14: Demographics of the Research Sample and of the Chinese Internet Population

	protester violent			protester improper			joining protest		
	college	age	gender	college	age	gender	college	age	gender
criminal	0.16	−0.14	−0.08	0.10	0.27	0.15	1.16***	0.51**	0.53**
	(0.14)	(0.14)	(0.15)	(0.19)	(0.20)	(0.21)	(0.21)	(0.20)	(0.22)
criminal:trust	−0.17	0.17	0.03	−0.01	−0.34	−0.17	−1.74***	−0.91***	−0.90***
	(0.23)	(0.22)	(0.23)	(0.31)	(0.30)	(0.31)	(0.33)	(0.31)	(0.32)
trust	−0.21	−0.12	−0.11	−0.05	−0.01	0.06	0.33	0.17	0.02
	(0.16)	(0.16)	(0.17)	(0.22)	(0.22)	(0.23)	(0.24)	(0.23)	(0.24)
Num. obs.	1328	1328	1328	1331	1331	1331	1331	1331	1331

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 15: Effect of Negative Labels (Government Source): Sample Reweighted