Supplementary material

Table of contents

Appendix A: Hypotheses 1
Research question 1: effect of an average disruptive protest 1
Research question 2: effect of protest/article characteristics 3
Appendix B: Power analysis 5
Pooled main effect 6
Characteristic effects 7
Appendix C: Article screening criteria and validation 11
Content criteria 11
Format criteria 12
Screening validation 13
Appendix D: LLM article summary validation 14
Appendix E: Attitude and background variable measurement 16
Appendix F: Article characteristic selection and coding procedure 17
Appendix G: Article characteristic coding pilots 20
Pilot 1: research assistants, objective characteristics 20
Pilot 2: all characteristics 24
Appendix H: Models 26
Effect of a typical media article 27
Effect by article characteristics 29
Multiple testing correction 30

Appendix A: Hypotheses

Research question 1: effect of an average disruptive protest

Our first research question is about the average effect of a media article about disruptive climate protest on environmental attitudes. We focus on five attitudes that are especially relevant to furthering climate policy and/or producing collective action on climate change. We hypothesize that exposure to a representative sample of articles about disruptive climate protest will on average:

H1. increase the salience of climate change;

H2. increase concern about climate change;

H3. increase support for climate change policy;

H4. increase intentions to take climate-related collective action;

H5. decrease identification with environmentalism.

The salience of climate change captures how much it is at the top of people's minds as a societal problem. Increasing climate salience through media attention is a typical stated goal of (disruptive) climate protest organisers, in order to put pressure on politicians. Indeed, public preferences are more likely to be reflected in policy outcomes when the issues at stake are highly salient (Giger and Lefkofridi, 2014; Lax and Phillips, 2012; Rasmussen et al., 2019).

Next, we look at concern about climate change–a common way of measuring general engagement with the issue (Bergquist et al., 2025).

We also measure support for climate policy and intentions to take collective action. These are perhaps the most politically relevant aspect of environmental attitudes, given that public support is the most important predictor of climate policy adoption (Yeganeh et al. 2020). Likewise, collective action has been argued to be "the most efficient method of achieving emission reductions" (Roser-Renouf et al., 2014, p.163; see also Schmitt et al., 2020). We note that if our items that measure collective action intentions had included joining protests, rather than the much more conventional behaviours that are in fact included, we might instead predict negative effects for collective action intentions. We also measure environmentalist identity, since it is proposed that de-identification from the movement and its issue is the main mechanism behind potential opinion backlash (Feinberg et al., 2020; Kenward, 2024; Simpson et al., 2022).

Lastly, we test more directly whether (loss of) environmentalist identity does indeed mediate the effect on other attitudes. This could be the case even if, as we hypothesize, main treatment effects on those attitudes are positive: a negative mediated effect could be outweighed by a positive direct effect. In order to limit the number of statistical comparisons, we create one compound environmental attitude index containing all outcomes except salience and identity. We hypothesize that:

H6. the effect of protests on environmental attitudes as mediated via environmentalist identity is negative;

H7. the direct, non-mediated effect of protests on attitudes is positive.

We only test these hypotheses if environmentalist identity is indeed affected by the protest treatments, as this is a necessary condition for such a mediation effect to occur.

Research question 2: effect of protest/article characteristics

The second research question of this paper is about the effect of protest and article *characteristics* on attitudes. As the outcome, we once again use the compound environmental attitude scale. We hypothesize that the effect of disruptive protest articles on attitudes depends on:

H8. the number of protesters;

H9-12. the level of disruption caused to businesses, the public, government, and culture/sports;

H13-16. whether blockades, attaching/climbing, event interruptions, or vandalism-like tactics are used;

H17. how much the messaging is also about non-environmental issues;

H18. the tone of the article;

H19-21. the amount of words the article spends on the protesters' message, on the disruption, and on negative comments;

H22. the total perceived amount of disruption caused;

H23. whether the protesters come across as ordinary people;

H24. whether the form of protest is judged to be acceptable.

The protest and article characteristics whose impacts we chose to evaluate are based on a combination of theory and existing findings. Large numbers could lend legitimacy to the protesters' demands and cause others to jump on the bandwagon (Tilly, 1995; Schmitt‐Beck, 2015)—though Wouters (2019) does not find evidence for this. The targets and tactics chosen for a disruptive protest are known to change its perceived appropriateness (Badullovich et al., 2024), which could in turn drive persuasion or backlash. Bringing in non-environmental issues could also hamper persuasion by alienating people who do not agree with the protesters on those issues (Marshall et al., 2024). The tone of the portrayal of protesters and the space given to criticism are a straightforward case of framing effects, as the media outlet and commenters are suggesting to the reader how to feel about the protest (McLeod, & Detenber, 1999; McLeod & Hertog, 1992).

If attention is given to the message, protesters not only have more opportunities to verbally persuade the reader–they are also given legitimacy (Brown & Mourão, 2021). Attention to the disruption, on the other hand, can be delegitimizing (id.). Protests that are perceived as very disruptive, and protesters who come across as very different from ordinary people, may also be especially likely to trigger de-identification from their cause (Feinberg et al,. 2020; Simpson et al., 2022). Finally, the perceived acceptability of a protest can be seen as the final product of all these factors. It may seem obvious that a protest seen as less acceptable will have more negative outcomes. However, given the number of studies that find no difference between the effect of disruptive and non-disruptive protests on attitudes towards the cause (Brehm and Gruhl, 2024; Bugden, 2020; Fuller et al. 2025; Gonzatti et al., 2023; Vandeweerdt, 2025), this is worth investigating further.

Appendix B: Power analysis

In this power analysis, we attempt to predict the statistical power of our experiment, both for the average treatment effect (across all articles) and for the differential effects of articles by their characteristics (e.g. whether the featured protest targeted ordinary citizens). Our analyses produce a set of power estimates, depending on the size of the expected treatment effect. Our power estimates are based on *p*-values corrected for multiple testing (see Appendix H).

We use an existing dataset from an experiment by Vandeweerdt (2025) measuring similar variables to the current study, in a comparable context. UK respondents (n=846) indicated their climate attitudes in a first wave; then half of respondents were exposed to a news article about a Just Stop Oil protester scattering orange powder paint on a snooker table during the World Snooker Championship. We adjust the treated group's outcomes so that the treatment effect is zero on all outcomes. This gives us a simulated pool of 846 untreated respondents. Relying on this data allows us to avoid making assumptions about the distributions of the outcomes, and how well they autocorrelate over time.[1]

Our power analysis approach comes down to sampling respondents with replacement from that experiment until we have 3325 simulated respondents (our planned sample size, n=3500 with a 95% retention rate). Then, we add various treatment effects of the articles. We run power analyses for both the pooled average treatment effects (ATE) of all the articles compared to the

[1] The latter dictates how much power we gain from controlling for pre-treatment outcomes. The time between waves in this previous study was two weeks, identical to the current one.

control group, and the ATE of each protest/article characteristic. We only investigate positive effects, because the results for negative effects would be symmetrical.

Pooled main effect

The original study has three outcomes of interest to us: concern about climate change, support for climate policy, and pro-environmental behavior intentions. All are made up of two-item scales going from 1 to 7. We run power analyses for detecting ATEs on all three outcomes. We calculate power levels for four different scenarios, from very small (.15) to moderate (.4) ATEs.

We proceed as follows:

1. Evenly adjust the outcomes of respondents in the original sample's treatment group so that the treatment effect becomes zero, in order to simulate a pool of 846 untreated respondents.
2. Repeat the following steps 250 times for each scenario and each outcome:
    a. Randomly allocate 80% of the simulated pool to the treated group.
    b. Increment the outcomes of respondents allocated to treatment, so that the true ATE corresponds to the scenario.
    c. Sample 3325 respondents with replacement from this pool and assign the treated group respondents to one of the 100 articles. Add a random effect of the article they were assigned to (drawn from a normal distribution, N with mean 0 and s.d. 0.1).
    d. Estimate the ATE using the planned analysis, clustering by article (see Appendix H). Store whether the effect was statistically significant.

3. To calculate power, for each outcome, average the significance result across all repetitions within the same scenario.

Table A.1 shows the resulting power levels for the different ATE scenarios. We see excellent power levels to detect even small effects (of just d=0.15) for all outcomes.

Table A.1

Simulated Power To Detect An Effect In Four ATE Scenarios, By Outcome.

ATE (Cohen's d)	**0.15**	**0.2**	**0.3**	**0.4**
Concern	0.992	1.000	1.000	1.000
Policy	1.000	1.000	1.000	1.000
Behavior	0.981	0.996	1.000	1.000

Characteristic effects

In addition to the average effect of all articles, we also calculate power for the effect of protest/article characteristics. For these analyses, we will use an outcome scale composed of all attitudinal items (except salience and environmentalist identity). Vandeweerdt's (2025) study contains two items about concern about climate change, one item about satisfaction with government action, support items for four climate policies, and behavioural intention items for two collective climate action behaviours. We pool all these items into a scale, using the same procedure as we plan on using for our own analyses. Power analyses for the differential effect of articles by their characteristics use this outcome scale.

For each of the characteristics, we calculate power levels under four scenarios, where those characteristics have small (.15) to large (.4) effects. For this exercise, all characteristics are rescaled to have 0 as their minimum and 1 as their maximum. In other words, the ATEs are the effect on environmental attitudes of moving from the lowest to the highest observed value on that characteristic. This means the effect sizes we investigate are conservative.

The joint distribution of characteristics across articles comes from a pilot study where we measured all characteristics across 50 articles (see Appendix G, section Pilot 2: all characteristics). As a result, we can take into account the fact that some characteristics have skewed distributions and some covary with each other–making their effects harder to detect.[2]

Note that the average treatment effect of reading an article (compared to the control group) is unimportant here, because the analysis does not involve the control group. Instead, it involves comparing the attitudes of respondents who read different articles.

We take the following steps:

1. For each respondent, pool their outcomes of interest into a scale.
2. Evenly adjust the outcomes of respondents in the original sample's treatment group so that the treatment effect becomes zero, in order to simulate a pool of 846 untreated respondents.
3. Get the real characteristic ratings of the articles that were included in Pilot 2, duplicating the entries to simulate a set of 100 articles.
4. Repeat the following steps 3000 times:
    a. Randomly choose one-third of the characteristics to have an effect, reflecting the realistic assumption that most characteristics will end up having null effects. Independently pick an effect size (ATE) scenario for each characteristic.

[2] An important caveat here is that, in order to protect our embargo on the articles to be used in the final study, this pilot was done on articles drawn from a two-year period preceding the actual article sampling period (see main text section "Search time period"). It is possible that some protest characteristics have become more or less rare since then.

b. Sample 2260 respondents with replacement from the simulated untreated pool; this is the expected number of treated respondents.
c. Randomly (but evenly) assign respondents to the articles. Then add treatment effects to respondents' environmental attitudes scale based on the characteristics of their assigned article.
d. Estimate the effect of each characteristic using the planned analysis script (see Appendix H). Store whether the effect on the outcome scale was statistically significant.

5. To calculate power, for each characteristic, average the significance result across all repetitions where it was assigned the same scenario (e.g. there will be about 250 repetitions where "general public targeted" was assigned an effect of 0.3–average over those).

Table A.2 shows power levels for the different characteristic effect scenarios. Power is rather different for different characteristics, even when they have the same expected effect size. This is because it is more difficult to detect the effect of a characteristic that is skewed (e.g. a binary one that is "on" in just 10% of the articles) or highly correlated with another characteristic.

All in all, though, power is acceptable even for relatively small effect sizes for all characteristics, with the exception of *Disruption to a cultural or sporting event or institution*. The reason for this is that it is strongly correlated with the use of *Interrupting an event* as a tactic (r=.57), making their independent effect difficult to discern. Cultural institutions are highly relevant to research, because they are seen as some of the most inappropriate targets for disruptive activism (Badullovich et al. 2024). It is not unthinkable that choosing such a target would have large (negative) effects. For that reason, we decided to retain the characteristic. We will appropriately qualify our conclusions about its effect with the level of uncertainty surrounding them.

Table A.2

Simulated Power To Detect The Effect Of Each Protest Characteristic In Four ATE Scenarios.

ATE (Cohen's d)	**0.15**	**0.2**	**0.3**	**0.4**
No. of protesters	0.776	0.888	0.996	1.000
Disruption targets				
Disruption to general public	0.637	0.790	0.957	0.996
Disruption to environmentally damaging businesses	0.609	0.698	0.890	0.983
Disruption to authorities	0.784	0.933	0.992	1.000
A cultural or sporting event or institution	0.445	0.647	0.935	1.000
Disruption tactics				
Blockade	0.983	1.000	1.000	1.000
Attaching/Climbing	0.992	1.000	1.000	1.000
Alterations/Vandalism	0.945	0.986	1.000	1.000
Interrupting an event	0.943	0.996	1.000	1.000
Issue bundling	0.879	0.967	1.000	1.000
Tone of portrayal	0.958	0.993	1.000	1.000
Article word counts (logged)				
on message	1.000	1.000	1.000	1.000
on disruption	0.996	1.000	1.000	1.000
on negative comments	0.890	0.967	1.000	1.000
Perceived overall disruption	0.738	0.819	0.943	1.000
Protesters seeming ordinary	1.000	1.000	1.000	1.000
Acceptability of protest	1.000	1.000	1.000	1.000

Note. ATE is the effect of changing the characteristic from its minimum to its maximum.

Appendix C: Article screening criteria and validation

Here, we describe the criteria that an article should meet in order to be filtered into the population of disruption articles. This is the population from which the final sample of treatment articles are selected (at random but weighted by the outlet's readership). There are both content eligibility criteria and format criteria. Next, we describe the steps we took to validate the performance of the LLM (Claude Sonnet 3.5) in screening the articles by these criteria.

We use the shorthand "disruptive climate protests" to refer to the types of protests we are interested in, because all disruptive environmental protest we identified during the test period foregrounded climate. However, if we identify disruptive environmental protest that does not mention climate, we will include it, and our LLM prompts use the term "environmental protest".

Content criteria

We evaluate the content eligibility for each article in two steps:

1. The article must first meet the broader criterion: "In a counterfactual world identical to ours, except that environmental protesters *never used disruptive methods*, this article could not exist in a form where the primary message was hardly changed." This criterion is assessed by Claude. We support this step by doing quality control of the final selected sample of treatment articles, flagging and swapping out any articles that do not in fact meet this criterion. Use of LLM facilitates our workflow despite the necessity for manual checking of screened in articles because we do not manually assess automatically screened out articles.
2. The article must then meet a narrower criterion: "in a counterfactual world where climate protesters had not used disruptive methods at *one or more specific protests*, the article

could not exist in a form where the message was hardly changed." To qualify, the article must mention these protests explicitly. Among the articles that pass step 1, we evaluate this criterion fully by hand.

The reason why we filter in two steps, rather than going directly to the narrow criterion in step 2, is because articles that pass step 1 but not step 2 are used in an additional analysis. This analysis is described in the main text (see section "Effect of all articles owing their existence to disruptive protests"). It is aimed at assessing the public opinion effects of disruptive protests more comprehensively.

Format criteria

In order to end up in the treatment population, the article must also be judged by Claude to pass the following format checks:

- It may not consist of one or more readers' letters to the editor. Letters tend to be short and include no headline or pictures, and they are not the same as debate articles, which tend to be longer and include those elements.
- It may not be a "news roundup", i.e. a collection of short updates or links to other articles. These updates are often thematically related but the article as a whole is not a coherent single article of the type that would appear in a print edition.
- It may not be in a language other than English (e.g., Welsh articles from the BBC) or clearly originate from a non-UK edition of the news outlet (for example, the Guardian and the Daily Mail both have Australian editions).
- It may not be centered almost exclusively around one or more videos, meaning there is no text in the article that is not video caption (serving the purpose of explaining the video).

In the quality control step, a human will flag and swap out any such articles that were incorrectly filtered in by Claude.

Screening validation

To validate the LLM's performance in applying these filtering criteria, we compared its screening decisions against a research assistant's screening decisions for 50 randomly selected articles (confusion matrix in Table C.1).

Table C.1

Confusion Matrix for Human vs LLM Screening Criteria Coding.

		LLM codes article as meeting criterion	
		No	Yes
Human codes article as meeting criterion	No	22	2
	Yes	1	25

The main quantity of interest here is recall; that is, the percentage of true positives (articles that should be screened in) that are caught as positives by the LLM (i.e. are indeed screened in). Recall is satisfactory at 96%, implying that only 4% of the articles that should have been in our stimulus pool were incorrectly filtered out.

The precision of the model–the rate at which the articles it filters in are true positives rather than false positives–is 93%. This means that among the filtered-in articles, 7% should not have been in the stimulus pool. This aspect of performance is less relevant, because we will manually evaluate each filtered-in article, deleting it from the stimulus sample if we judge it to be a false positive. Deleted articles are replaced by another randomly selected article from the same outlet.

Appendix D: LLM article summary validation

If the original article is longer than 350 words, we ask Claude to summarize it down to 300. In a pilot testing (see Appendix G, pilot 2), this was necessary for 43 out of the 50 articles that passed screening.

Based on an assessment of approximately 200 summaries (drawn from a two-year period before the actual research period, see main text section “Article search”) we judged the quality and fidelity of these summaries to be high, with the caveat that some minor mistakes would have to fixed by hand in a quality control stage. For instance, despite being instructed to the contrary, Claude occasionally included meta-text such as “the article states that….”.

At an early stage of piloting the article summary process, we prompted the LLM to code original articles, and summarised versions of them, for a range of article characteristics similar to the final characteristics we have now selected. We found no evidence that the summary process systematically influenced the coded article characteristics, so our planned validation is likely to be successful. However, because this early analysis was not based on the exact finally selected article characteristics, and did not involve a human coder, we do not present these pilot results in detail.

Our validation will start from the list of characteristics that we believe might have an impact on attitudes (see subsection “Protest and article characteristics” in the main text). These include characteristics of both the protest and its presentation in the article. The validation procedure will show us whether the summarisation process preserves the distribution of characteristics. This is key in order to preserve the representativeness (external validity) of our average treatment effect. If the summarization systematically changed the ratings in a particular direction (e.g. making the tone more neutral despite our prompt instructions to preserve the tone), it may introduce bias that would skew the average treatment effect in a particular direction.

We will ask a research assistant (RA) to first code the summaries of a randomly selected subset of 40 articles on all characteristics that are to be rated by RAs in the final study–namely, all objective features of the protest plus the article's tone in portraying the protest. At least one month later, we will ask them to rate the full-length version of the articles. We will then compare average ratings on the characteristics between the full-length and summarised versions (using Bayes factors, as we are trying to prove the null of no difference). With respect to the article's tone in portraying, the bias is likely to take the form of articles having a less clear tone (whether negative or positive). For that reason, we will also investigate a version of the variable that is "folded" in the middle (i.e. turning the five-point scale into neutral tone, mild tone, strong tone). This way, we can see whether the intensity of the tone decreased.

If, for any characteristic, the validation shows that summarization introduces a systematic bias, then we will proceed as follows. First, we will investigate whether an LLM can be used to efficiently spot articles where the characteristic changed due to the summarization. That is, we will check whether an LLM can be used to replicate the RA's judgments. If so, we will use the LLM to compare the original version and summary of all articles in the final stimulus set of 100, flagging articles for which it would rate the original and the summary differently. Next, we will manually re-adjust the characteristic in the flagged articles. For the objective protest characteristics, we will use our own judgment to ensure that the adjusted summary contains the same information as the original. For the article's tone, we will once again use an LLM to check whether the manual adjustment would now result in identical ratings between the original and summary. If the initial tests show that it is not possible to replicate human judgments in this way, then we will adapt this process to use human raters instead.

Appendix E: Attitude and background variable measurement

Table D.1 documents the response scales and the sources of the items used to measure environmental attitudes in the survey experiment. Table D2. shows the background variables we collect, and their response options.

Table D.1

Environmental Attitude Measurement Details.

Outcome	Response scale	Notes
Salience	Coded as 1 if respondent mentions "climate", "biodiversity", "pollution", "environment", "nature", or "warming"	Standard "most important problem" item; used as an outcome of disruptive protest by Vandeweerdt (2025).
Climate change concern	Strongly Disagree (1), Disagree (2), Somewhat disagree (3), Neither agree nor disagree (4), Somewhat agree (5), Agree (6), Strongly agree (7)	From Kenward and Brick (2024).
Climate policy support	Strongly Disagree (1), Disagree (2), Somewhat disagree (3), Neither agree nor disagree (4), Somewhat agree (5), Agree (6), Strongly agree (7)	From Akehurst (2024), based on YouGov surveys
Collective action intention	Very unlikely (1), Unlikely (2), Somewhat unlikely (3), Neither likely nor unlikely (4), Somewhat likely (5), Likely (6), Very likely (7)	From the participatory actions subscale of the validated Environmental Action Scale by Alisat & Riemer (2015), updated by Oinonen and Paloniemi (2023) to reflect new online tools.
Environmentalist identity	Strongly Disagree (1), Disagree (2), Somewhat disagree (3), Neither agree nor disagree (4), Somewhat agree (5), Agree (6), Strongly agree (7)	From Brick and Lai (2018).

Table D.2

Background Variable Measurement Details.

Outcome	Response scale	Notes
Sex	Male (1), Female (2)	Provided by Prolific
Age	In years	Provided by Prolific
Region	North East, England (1), North West, England (2), Yorkshire and the Humber, England (3), East Midlands, England (4), West Midlands, England (5), East of England 6), London, England (7), South East, England (8), South West, England (9), Wales (10), Scotland (11), Northern Ireland (12)	Provided by Prolific
Education	No formal qualifications (1), Secondary education (e.g. GED/GCSE) (2), High school diploma/A-levels (3), Technical/community college (4), Undergraduate degree (BA/BSc/other) (5), Graduate degree (MA/MSc/MPhil/other) (6), Doctorate degree (PhD/other) (7)	Provided by Prolific
Ideology	Very left-wing (1), Fairly left-wing (2), Slightly Left-of-centre (3), Centre, (4) Slightly right-of-centre (5), Fairly Right-wing (6), Very right-wing (7), Don't know (99)	From Newman et al. (2024)

Appendix F: Article characteristic selection and coding procedure

Each news article stimulus must be rated on a number of characteristics that might influence its effect on environmental attitudes. We subdivide these characteristics into “objective” traits, which will be rated by two research assistants (RAs); traits that are encoded in the count of words in the article dedicated to different aspects (e.g. consequences of the disruption), which will be rated by an LLM (Claude Sonnet 4.0); and "subjective" traits, which will be rated by a panel of 14 crowd-coders for each article.

Table E.1 below contains the wordings of the questions that we will ask of the research assistants and crowd coders. Each crowd-coder will have 20 minutes to read instructions and then code five summarized articles (pilot studies showed that this timing is reasonable). For characteristics marked with an †, coders will be given the opportunity to indicate that the article does not contain enough information to form a judgment. If they do so, their rating will not be used. If at least half of the coders judge this to be the case for a given characteristic of an article, then that characteristic of that article will be considered missing.

An analysis of data from Badullovich et al. (2024) suggested that ideology is by far the most important driver of people's opinions of disruptive protests, and that other demographics (age, gender, education) have little explanatory power when ideology is controlled for. We therefore ensure that (as far as possible) every article is read by a set of 14 coders that is balanced on ideology.

Table E.1

Characteristic Coding Scheme.

Characteristic	Raters	Question
No. of protesters†	RA	What is your best estimate of how many people are taking part in the protest? (8-point scale, e.g. "4 to 10", "11 to 30", "31 to 100")
Disruption targets	RA	How much disruption does the article describe as being caused by the protest to each of the following entities? (5-point scale from None at all to A great deal)
Disruption to general public		The general public going about everyday activities
Disruption to environmentally damaging		Business operations usually perceived as environmentally damaging

businesses		
Disruption to authorities		Government or local authority operations (not including police)
Disruption to culture/sport		A cultural or sporting event or institution
Disruption tactics	RA	What types of disruption are mentioned as being part of the protest? (binary, multiple selections allowed)
Blockade		Blockade
Attaching/Climbing		Protesters physically attached to or on top of something
Alterations/Vandalism		Temporary physical alterations or vandalism
Interruption		Interrupting an event or speech
Issue bundling	RA	To what extent do the protesters mention other political issues in their messaging that go beyond the environment? (5-point scale from Not at all to Far beyond environmental issues)
Tone of portrayal	RA	Would you say that the article portrays the protest in a hostile, neutral or sympathetic way? (5-point scale from Very hostile to very sympathetic)
Article word counts	LLM	Prompt asking LLM to tag all words and pictures in the article that cover…
on message		…the messages protesters are trying to convey, for example quotes (from a person or a banner) or reported summaries of things protesters said.
on disruption		…general disruption effects (direct or knock-on) of the disruptive protest
on negative comments		…criticism of the protest or its effects. This might be expressed by a quoted individual, reported in general terms (e.g., "passers-by shouted disapprovingly"), or expressed by the author of the article.
Perceived overall disruption†	Crowd	How much disruption would you say the protest caused? (5-point scale from None at all to A great deal)
Protesters seeming ordinary†	Crowd	Besides being environmental activists, would you say the protesters come across like ordinary people? (5-point scale from Not at all to Very much like ordinary people)
Acceptability of	Crowd	Do you find this form of protest acceptable or unacceptable?

protest	(5-point bipolar scale from Very unacceptable to Very acceptable)

For the LLM-based word counts, we will also take into account pictures that convey either the message of the protesters (e.g. a photo of a banner) or the disruptive effects of the protests (e.g footage of queuing cars). These are a key way for news outlets to frame protests (Brown & Mourão, 2021). We will ask Claude to code whether each picture was about either of those aspects, and count a picture as being equivalent to 50 words. The median summarized article in our pilot test (see Appendix G, Pilot 2) contained 315 words and three pictures. All words and pictures in the article can in principle be assigned to more than one category at once. We take the natural logarithm of the picture-augmented word counts (plus one) before adding them to the model, both because we believe that extra words will have decreasing marginal effects, and to limit the influence of outliers.

Appendix G: Article characteristic coding pilots

We piloted the article/protest characteristic coding scheme twice: once using research assistants (RAs), and then again using the combination of rater types that we plan to use in the final study (RAs, crowd coders, and the LLM Claude). Both pilots used articles drawn from a two-year period that preceded the actual article sampling period (see main text section “Article search”).

Pilot 1: research assistants, objective characteristics

In the first pilot, we had two research assistants rate a total of 50 randomly selected summarised articles. They rated a broad range of “objective” characteristics of the protests. The purpose of this pilot was to calculate the distribution and the inter-rater reliability for these objective characteristics. This way, we could identify any characteristics that either (1) were too rare or skewed to be worth including in an analysis, or (2) could not be reliably measured from the information that was present in a typical summary.[3]

Table F.1 shows the results of this initial pilot. As a measure of inter-rater reliability, we report the intra-class correlation (ICC) for the average rating. Specifically, we use the two-way mixed effects model, because the same two raters scored all articles, so any bias due to the rater will affect all articles equally.[4] Koo and Li (2016) interpret ICC values as poor if less than 0.5, moderate if between 0.5 and 0.75, good if between 0.75 and 0.9, and excellent if greater than 0.90.

Table F.1

First Coding Scheme Pilot Results.

Characteristic	Mean	ICC (3,k)	Decision
Number of protesters (9-point scale with intervals, e.g. “3 to 5”)			
Number of protesters	5.2	.95	Adjust interval options
Share of protestors wearing costumes (scale from 1 - None to 4 - Most or all of them)			
Share of protestors wearing costumes or masks	1.6	.64	Drop from study, 62% missing data
Level of disruption, by target (scale from 1 - None at all to 5 - A great deal)			
Business operations usually perceived as environmentally damaging	1.9	.43	Keep as is but develop additional coder guidance
Business operations not usually perceived as environmentally damaging	1.6	.46	Keep as is but develop additional coder guidance

[3] In addition, the RAs rated the articles on the subjective characteristics that would ultimately be judged by crowd coders. We did this in order to fine-tune the wording of the questions asked of crowd-coders, based on the rater’s qualitative feedback.
[4] In other words, we care that raters should be internally consistent, not that their scores should be correct in an absolute sense. This is also known as the ICC (3,k) according to the Shrout and Fleiss (1979) convention.

One or more specific individuals perceived as responsible for environmental damage	1.1	-.10	Drop from study, low ICC
Government or local authority operations	1.5	.44	Keep as is but develop additional coder guidance
The general public going about their everyday lives	2.2	.47	Keep as is but develop additional coder guidance
The general public going about specific activities perceived as environmentally damaging	1.5	-.02	Drop from study, low ICC
Caring-related activities (medical, care of children or elderly, funeral, etcetera)	1.1	-.06	Drop from study, low ICC
A cultural or sporting event or institution	1.2	.86	Keep as is
Disruption tactics used (binary, not mutually exclusive)			
Blockade	73%	.78	Keep as is
Attaching/Climbing	51%	.65	Keep as is
Temporary physical alterations	6%	.46	Join with vandalism
Vandalism	7%	.70	Join with temporary alterations
Interrupting an event or speech	12%	.77	Keep as is
Breaching a legal injunction	6%	.79	Drop from study, too rare
Protester demographics mentioned (binary, not mutually exclusive)			
Student	26%	.51	Keep as is
Unemployed	2%	.00	Drop from study, too rare
Having the type of occupation that ordinary people do not have	2%	-.04	Drop from study, too rare
Very young	23%	.21	Drop from study, low ICC
Very old	17%	.46	Drop from study, too rare
Private-school educated	1%	.00	Drop from study, too rare
Radicalness of the demand (scale from 1 - Very radical to 5 - Very moderate)			
Radicalness	2.8	.03	Drop from study, low ICC

Messaging on issues beyond environment (scale from 1 - Not at all to 5 - Far beyond)			
Issue bundling	1.91	.73	Keep as is; the actual study period will likely contain more of this
Portrayal of protesters in article (scale from 1 - Very hostile to 5 - Very sympathetic)			
Tone of portrayal	2.7	.59	Add examples to each level

Note. Ratings by two research assistants. Means are across all articles and raters.

In some cases, the pilot revealed that the characteristic did not occur often enough, or could not be measured reliably enough, to make an analysis of their effect feasible. For example, most articles did not include enough information for both raters to score it on the *Share of protestors wearing costumes or masks* (and when they said masks were present, they were largely incorrectly counting Covid masks). Similarly, we found that there were zero articles in which both RAs agreed that there was *Disruption to a caring-related activity*. Preliminary power analyses confirmed that the effect of such variables could not be reliably detected. These characteristics were therefore dropped from the study (as noted in table F.1, last column).

Two exceptions to this are *Disruption to Government or Authority* and *Disruption to a cultural or sporting event or institution*, which we decided to keep in the study despite being relatively rare (and, in case of the former, difficult to measure). One reason is that they are theoretically important. Badullovich et al. (2024) find that government officials are seen as a highly acceptable target for climate-related civil disobedience, second only to fossil fuel companies. Museums on the other hand are seen as the most unacceptable target (along with ordinary people). So, we found it worthwhile trying to detect whether these choices of targets change the public opinion effects of protests. A second reason is that simulations showed that despite their rarity, their effect could still be detected at acceptive levels of power (see Appendix

B, Table A.2). This is because they have low correlations with the other variables that are modeled alongside, meaning that their independent connection with public opinion is relatively easy to establish.

Finally, the responses from the RAs made it clear that some adjustments to the question wording were needed. For example, when rating how much disruption was caused for governments or local authorities, pilot coders occasionally included disruption of police functioning. However, we consider the police to be in a different category than government here: protesters tend to target governmental bodies intentionally and explicitly due to the power they have over climate policy, whereas any disruption to police is typically "collateral damage".

Pilot 2: all characteristics

The main purpose of this pilot was once again to identify any subjective characteristics that were too rare to be worth including in an analysis, as well as to evaluate the level of disagreement between crowd coders. This second pilot also produced the ratings that were used in our power calculations (see Appendix B), so that those calculations might be based on realistic assessments of how often different protest characteristics occur, and co-occur.

We randomly drew and summarized another 50 articles (after filtering to make sure they met the inclusion criteria, see Appendix C). Here, the characteristics were rated by a research assistant (for objective characteristics), three crowd-coders each (for subjective characteristics), and Claude Sonnet 3.5 (for word counts). The crowd-coders were UK residents recruited on Prolific (n=30) and rated five articles each, resulting in three coders per article.

Table F.2 summarizes the results of this second pilot, showing mean ratings as well as the decision to include or exclude the tested characteristics from the final study. We also report the ICC for the crowd-sourced characteristics. We once again use the ICC of the average rating, but this time with random effects of rater, because each article gets a random set of raters.[5] Note that in the final study, the ICCs of the average ratings will be significantly stronger than the ones reported here, because we will go from three to 14 raters.

Table F.2

Second Coding Scheme Pilot Results

Characteristic	**Mean (or % present)**	**ICC (2,k)**
RA-rated characteristics		
Number of protesters (9-point scale with intervals, e.g. "4 to 10")	3.88 (median "11 to 30")	-
Level of disruption, by target (scale from 1 - None at all to 5 - A great deal)		
Business operations usually perceived as environmentally damaging	1.72	-
Business operations not usually perceived as environmentally damaging	1.18	-
The general public going about their everyday lives	2.48	-
Government or authorities	1.40	-
A cultural or sporting event or institution	1.42	-
Disruption tactics used (binary)		
Blockade	76%	-
Protesters attached to something	58%	-
Temporary alterations or vandalism	22%	-
Interrupting an event or speech	16%	-
Demographic, student	10%	-
Tone of portrayal (scale from 1 - Very hostile to 5 - Very sympathetic)	3.06	-
LLM-rated (article word count) characteristics		

[5] This also means that now, we do care about raters being correct in an absolute sense, since we do not want an article's score to depend on the subset of raters that it happened to be assigned to. This version of the ICC is also known as the ICC (2,k) according to the Shrout and Fleiss (1979) convention.

Words on message	93	-
Words on disruption	209	-
Words on negative comments	36	-
Words on positive comments	2	-
Crowd-sourced characteristics		
Perceived overall disruption (5-point scale)	3.50	.94
Protesters seeming ordinary (5-point scale)	3.07	.69
Acceptability of protest (5-point scale)	2.13	.81

Note. ratings by one RA, three crowd-coders per article, and Claude.

Based on these results, we decided to drop three further characteristics. First, we deleted *Disruption to business operations not usually perceived as environmentally damaging* because in this pilot, it was much rarer (only 12% of cases having a value above "1 - None"). As a result, power calculations for the effect of this characteristics are not convincing. We deleted *Demographic, student* for the same reason. Lastly, we dropped the LLM-rated words spent on positive comments. Positive comments were exceedingly rare and when they occurred, they were extremely brief.

In our description of the results of the experiment, when we report on analyses involving characteristics with an ICC of 0.7 or lower (as calculated on our final set of treatment articles, using 2 RAs and 14 crowd-raters per article), we will briefly note their potentially less reliable measurement.

Appendix H: Models

All responses complete for both waves will be included in analysis; partial responses will be left out of analyses for which they are missing the required variables. Below, we provide pseudo-code and details for each of the models used in our analyses.

Effect of a typical media article

In this analysis, we estimate one model per attitudinal outcome. Written in pseudo-R code, this model is:

fit <- lm(outcome_wave2 ~ treated + outcome_wave1, data=d)

The quantity of interest is the coefficient of *treated*. Although it is typical practice within the stimulus sampling approach, we do not model article identity as a random factor. We conducted simulations indicating a heavy reduction in power to detect a treatment effect, likely because there is only one level of control, so the random component of the model absorbs much of the treatment vs. control difference. We therefore take an alternative approach to modelling the relations between participants assigned to the same article: clustered standard errors. Like random factors, these ensure that our uncertainty estimates take into account the fact that the stimuli are a sample of, rather than the full population of, all possible stimuli representing our treatment concept (Judd et al., 2012; McNeish et al., 2017). We will calculate standard errors as follows, using the *lmtest* and *sandwich* R packages:

coeftest(fit, vcov = vcovCL(fit, cluster = ~article))

The model for the alternative analysis, where observations are weighted by source and ideology, is:[6]

dat.weighted <- svydesign(ids = ~article, data = d, weights = prob_reading)

svyglm(outcome_wave2 ~ treated + outcome_wave1, design=dat.weighted)

where *prob_reading* is the probability that the tested combination of respondent-ideology and article would occur in the real world, based on the Reuters Digital News Report survey (Newman et al. 2024).

[6] *svydesign* and *svyglm* are functions adapted to deal with sampling weights. They are used to adjust for a known difference between respondent groups' (here: combinations of respondent ideology and news source) likelihood of ending up in the sample, and their frequency in the population.

To make this analysis possible, Wave 1 of our survey experiment includes the same seven-point ideology scale as the Reuters survey. However, we will collapse the two most extreme categories on either side. This is because in the Reuters study, there are not enough respondents in the most extreme ideological categories to accurately estimate their news consumption habits.[7]

Finally, we will perform an analysis to investigate whether the effect of the treatment on attitudes is mediated by its effect on environmentalist identity. We will only do this if the treatment has a statistically significant effect on identity. We pool all attitudinal outcomes (except salience and identity) into a compound environmental attitude scale, and then estimate both the identity-mediated and the direct effect of the treatment on that scale.

To do so, we start by taking the first difference (Wave 2 minus Wave 1 values) for both the mediator (identity) and the outcome (attitude scale), so we can test whether a *change* in identity due to the treatment mediates a *change* in the outcome. See Analyses subsection Mediation by environmentalist identity in the main text for a discussion of why this strategy helps with identification of the mediation effect. Next, we regress change in the mediator on the treatment variable; we regress change in the outcome on the treatment and change in the mediator; and finally we use the *mediation* package in R to use the estimates from these models to compute mediation effects. The pseudo-code is:

```
fit_MtoT <- lm(identity_change ~ treated, data=d)
fit_YtoM <- lm(outcome_scale_change ~ treated + identity_change, data=d)
mediate(fit_MtoT, fit_YtoM, sims=50, treat="treated", mediator="identity_change",
cluster=d$article)
```

[7] Namely, n=93 on the extreme left and n=36 on the extreme right. This means that for some of the less popular outlets, only two to four respondents in these categories say they consume the source.

The effect sizes of interest are (1) the average mediated causal effect of protests on attitudes via environmentalist identity, and (2) the direct effect of the protests when identity is held constant, a.k.a. the direct natural effect (Imai et al., 2011).

Effect by article characteristics

Here, we once again pool all outcomes (except salience and identity) into a compound scale. We then run four models, in which we progressively add variables, block by block. So, for instance, the first model will be:

lm(outcome_scale_wave2 ~ numberOfProtesters + ... + issue_bundling + outcome_scale_wave1, data=d[d$treated,])

Where … stands for the other characteristics in the first block (objective protest features). The final model will include all listed characteristics from all blocks:

lm(outcome_scale_wave2 ~ numberOfProtesters + ... + acceptability + outcome_scale_wave1, data=d[d$treated,])

Where … stands for the other characteristics in all blocks. We will cluster standard errors in the same way as in the main analyses of the average article treatment effect, using *lmtest* and *sandwich.*

The quantities of interest are the coefficients on the characteristics of the block that was just added to the model. These coefficients show how climate change attitudes among treated respondents vary depending on each characteristic of the protest or article, assuming the causal relationships between characteristics that are shown in Figure 1 (main text). Characteristics measured on Likert scales are modeled as continuous.

Three characteristics (*Number of Protesters*, *Perceived Overall Disruption and Protesters Seeming Ordinary*) can have missing values for some articles. This happens when most of the coders judge that there was not enough detail in the article to rate the characteristic (see Appendix F).[8] Still, we want to avoid dropping those articles from the analyses entirely. Instead, we will impute the median value for that characteristic in that article. We will also add a dummy variable to the analysis, which equals 1 when a respondent saw an article whose value on that characteristic was imputed. This keeps the characteristics' imputed values from interfering with their estimated effect.

This approach is plausibly unbiased for cases where the missing value is actually undefined (Allison, 2010)—as opposed to a respondent e.g. forgetting or declining to fill out a survey item. That is most likely the case for our missing values. When most raters found a characteristic to be unratable for an article—for example, when asked to rate the number of participants at a protest that had been announced but had not taken place yet—then that article simply has no value for that characteristic.

Multiple testing correction

Besides presenting the raw *p*-values, we will also show *p*-values corrected for multiple testing. They are calculated using the adaptive two-stage approach from Benjamini et al. (2006), which controls the so-called false discovery rate or FDR. The FDR is the rate at which discoveries (null hypothesis rejections) can be expected to be false. By setting the FDR to .05 we ensure that out of all confirmed hypotheses, the expected rate of false positives is no more than 5%. The approach adjusts p-values so that the FDR is indeed limited at .05.

We apply the approach separately to our two research questions, about (1) the average effect of exposure to an article on different environmental attitudes (H1-7), and (2) the effect of protest/article characteristics (H8-24). In other words, we control the FDR for the hypotheses

[8] In our (second) pilot study, this was the case for 0% of articles w.r.t. *Number of Protesters*, 4% of articles w.r.t. *Perceived Overall Disruption*, and 18% w.r.t *Protesters Seeming Ordinary.*

about the effect of reading a disruptive protest article on various environmental attitudes, and separately for the hypotheses about the effect of various protest characteristics on the aggregated environmental attitude scale. Since these are conceptually different questions, which speak to different theories and readers, we believe this is more meaningful than controlling for the FDR of all hypotheses combined. Our answer to question (1) should not be influenced by how many characteristics we decided to test for question (2), and vice versa for the number of attitudes tested in question (1).

We present both unadjusted and adjusted p-values because limiting the FDR has the obvious downside that it decreases power to detect true positives. Practitioners may care just as much about a false negative (e.g. we incorrectly conclude that the target of the disruption does not matter for public opinion) as a false positive (e.g. we incorrectly conclude that it does). The number of other tests run would not change this logic. We present the unadjusted p-values so that it remains possible to evaluate the evidence for each hypothesis separately, regardless of which others were tested.

References for supplementary material

Alisat, S., & Riemer, M. (2015). The environmental action scale: Development and psychometric evaluation. *Journal of Environmental Psychology*, *43*, 13–23.

Akehurst, S. (2024, September 27). What is living and what is dead in Europe's climate consensus? *Strong Message Here*. https://strongmessagehere.substack.com/p/what-is-living-and-what-is-dead-in (Retrieved July 5, 2025)

Allison, P. D. (2010). Missing data. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 631–657). Emerald Publishing.

Badullovich, N., Tucker, D., Amoako, R., Ansah, P., Davis, B., Horoszko, U., Zakiyyah, L., & Maibach, E. (2024). How does public perception of climate protest influence support for climate action? *npj Climate Action*, *3*(1), 16.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507.

Bergquist, P., Vandeweerdt, C., Mildenberger, M., Howe, P., & Marlon, J. (2025). *Global geographic variation in climate concern at national and sub-national scales*. [Manuscript Submitted for Publication.] https://osf.io/preprints/osf/s58ae_v1

Brehm, J., & Gruhl, H. (2024). Increase in concerns about climate change following climate strikes and civil disobedience in Germany. *Nature Communications*, *15*(1), 2916.

Brick, C., & Lai, C. K. (2018). Explicit (but not implicit) environmentalist identity predicts pro-environmental behavior and policy preferences. *Journal of Environmental Psychology*, *58*, 8–17.

Brown, D. K., & Mourão, R. R. (2021). Protest coverage matters: How media framing and visual communication affects support for Black civil rights protests. *Mass Communication and Society*, *24*(4), 576–596.

Bugden, D. (2020). Does climate protest work? Partisanship, protest, and sentiment pools. *Socius*, *6*, 2378023120925949.

Feinberg, M., Willer, R., & Kovacheff, C. (2020). The activist's dilemma: Extreme protest actions reduce popular support for social movements. *Journal of Personality and Social Psychology*, *119*(5), 1086–1105.

Giger, N., & Lefkofridi, Z. (2014). Salience-based congruence between parties and their voters: The Swiss case. *Swiss Political Science Review*, *20*(2), 287–304.

Gonzatti, D. S., Hunger, S., & Hutter, S. (2023). *Analysebericht zur Studie 'Environmental Protest Effects on Public Opinion: Experimental Evidence from Germany'* [Working paper]. https://doi.org/10.31219/osf.io/5mb3u

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology, 103*(1), 54.

Kenward, B. (2024). *Återställ Våtmarker (ÅV) impact evaluation - Preliminary report 0.6* [Restore Wetlands (RW) impact evaluation - Preliminary report 0.6]. https://www.benkenward.com/articles/preliminary_report0.6.pdf

Kenward, B., & Brick, C. (2024). Large-scale disruptive activism strengthened environmental attitudes in the United Kingdom. *Global Environmental Psychology*, *2*, 1–35.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Lax, J. R., & Phillips, J. H. (2012). The democratic deficit in the states. *American Journal of Political Science*, *56*(1), 148–166.

Marshall, R., Anderson, S. E., Van Boven, L., Al-Shawaf, L., & Burgess, M. G. (2024). Neutral and negative effects of policy bundling on support for decarbonization. *Climatic Change*, *177*(4), 61.

McLeod, D. M., & Detenber, B. H. (1999). Framing effects of television news coverage of social protest. *Journal of Communication*, *49*(3), 3–23.

McLeod, D. M., & Hertog, J. K. (1992). The manufacture of public opinion by reporters: Informal cues for public perceptions of protest groups. *Discourse & Society*, *3*(3), 259–275.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114–140.

Oinonen, I., & Paloniemi, R. (2023). Understanding and measuring young people's sustainability actions. *Journal of Environmental Psychology*, *91*, 102124.

Rasmussen, A., Reher, S., & Toshkov, D. (2019). The opinion-policy nexus in Europe and the role of political institutions. *European Journal of Political Research*, *58*(2), 412–434.

Roser-Renouf, C., Maibach, E. W., Leiserowitz, A., & Zhao, X. (2014). The genesis of climate change activism: From key beliefs to political action. *Climatic Change*, *125*, 163–178.

Schmitt, M. T., Neufeld, S. D., Mackay, C. M. L., & Dys-Steenbergen, O. (2020). The perils of explaining climate inaction in terms of psychological barriers. *Journal of Social Issues*, *76*(1), 123–135.

Schmitt-Beck, R. (2015). Bandwagon effect. In *The international encyclopedia of political communication* (pp. 1–5). https://doi.org/10.1002/9781118541555.wbiepc015 (Retrieved July 5, 2025)

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Simpson, B., Willer, R., & Feinberg, M. (2022). Radical flanks of social movements can increase support for moderate factions. *PNAS Nexus*, *1*(3), pgac110.

Tilly, C. (1995). *Popular contention in Great Britain, 1758–1834*. Harvard University Press.

Vandeweerdt, C. (2025). *The activist's trade-off: Climate disruption buys salience at a cost* [Manuscript submitted for publication]. https://www.claravdw.com/wp-content/uploads/2024/08/Activists-tradeoff.pdf

Wouters, R. (2019). The persuasive power of protest: How protest wins public support. *Social Forces*, *98*(1), 403–426.

Yeganeh, A. J., McCoy, A. P., & Schenk, T. (2020). Determinants of climate change policy adoption: A meta-analysis. *Urban Climate*, *31*, 100547.