Online Appendix for

Endogenous benchmarking and government accountability: Experimental evidence from the COVID-19 pandemic

British Journal of Political Science

Michael Becher (IE University)
Sylvain Brouard (Sciences Po)
Daniel Stegmueller (Duke University)

A. Appendix

A.1. Survey details 3
A.2. Pre-registration 4
A.3. Experiment 1 9
A.3.1. Vignette wording 9
A.3.2. Wording of key survey variables 10
A.3.3. Descriptive statistics of central variables 12
A.3.4. Respondent evaluations of experiment 12
A.3.5. Additional analysis of endogenous benchmark choice 14
A.3.6. Estimates of exogenous benchmark effect 14
A.3.7. Treatment effect heterogeneity 16
A.3.8. Impact of country references in vignette headlines 18
A.3.9. Benchmarks, performance evaluations, and vote choice 18
A.3.10. Additional experiment: Austria 21
A.4. Experiment 2 25
A.4.1. Vignette wording 25
A.4.2. Additional results 26
A.4.3. Additional analysis of endogenous benchmark choice 27

A.1. Survey details

Table A.1 provides fieldwork dates, sample size, response and completion rates by country for the survey in which we implemented experiment 1.

Table A.1
Survey details

	Fieldwork	Sample size	Resp. rate[a]	Completion rate
France	04/15 - 04/16	2 020	0.47	0.96
Germany	04/16 - 04/18	2 000	0.31	0.93
United Kingdom	04/15 - 04/17	1 000	0.35	0.94

a Response rate S/I, completion rate $C/(S-Q)$; I is the number of individuals invited, S the number of started surveys, Q number of surveys removed due to quota being fulfilled, C number of completed surveys.

This study, including experiment 1 and experiment 2, adheres to the *American Political Science Association*'s Principles and Guidance for Human Subjects Research and received IRB approval. The opt-in survey was conducted by Ipsos, a commercial polling company. The study does not include vulnerable groups or entail any physical or otherwise harmful interventions. Respondents are adults who have given their prior consent to be contacted to participate in a survey. Invitations to participate in our survey are emailed to the company's pool of respondents so that that share of respondents matches relevant quotas on the population margins with respect to variables like age, occupation and region of residence (quota sampling). Individuals choosing to opt-in to participate in the survey (on their computer or mobile phone) have to give their explicit consent. First, at the beginning of the survey, respondents must agree by reading the documents regarding data confidentiality and privacy policy and take an active action to give the consent (tick a special box stating "Yes, I agree"). Second, the survey informs respondents about the type of questions they will encounter in the survey and asks them for their informed consent. The survey covers questions about politics and political preferences, which may be seen as sensitive. However, we consider the risk as minimal because all countries are established democracies where opt-in surveys of this nature are common (e.g., European Social Survey, national election surveys).

A.2. Pre-registration

Both experiments were preregistered with the University of Pennsylvania-Wharton School's Credibility Lab. Both pre-analysis plans are included at the end of this section.The anonymized copy of the pre-analysis plan for experiment 1 can be retrieved at this this link: https://aspredicted.org/blind.php?x=8n2n54, and the anonymized copy of the pre-analysis plan for experiment 2 is available at this link https://aspredicted.org/blind.php?x=p4k2iu.

Below we summarize the mapping between the planned analysis in the pre-registration and results presented in the paper for each outcome variable. We also note any deviations from the plan.

- Dependent variable 2 (choice of benchmark text). Main results are in Figure 1 of the article. The pre-registered analysis uses pre-treatment satisfaction with the government as the main predictor (left panel). For similar results based on semi-parametric models that allow for a more flexible assessment of the relationship between pre-treatment satisfaction and headline choice, see Online Appendix Figure A.3. Following a reviewer suggestion, an additional analysis (right panel of Figure 1) uses party identification as a predictor. Results for Austria (different experimental design) are in Figure A.5.

- Dependent variable 1a (how government has handled the crisis compared to most other countries, abbreviated as COMPGOV). The main results are presented in Figure 2 of the article, without and with pre-treatment covariates. Covariate adjustment was not explicitly mentioned in pre-registration and is added as a robustness check. Additional results illustrating the effect size are summarized in Table A.3. Table A.3.7 reports the results of the pre-registered heterogeneity analysis. Results for the separate experiment fielded in Austria (there is no unconditional exogenous treatment, as noted in pre-registration and in Online Appendix A.3.10 below) are presented in Figure A.5.

- Dependent variable 1b (vote intention). Figure A.4 displays results based on a standard vote intention question (Measure 2 in pre-registration). In addition to the the analysis leveraging experimental variation in exogenous information, we also show results from an observational analysis of the correlation between COMPGOV and vote intention. Note that the pre-registration also includes another vote intention variable (Measure 1). However, this variable had to be dropped from the survey in France and UK before field work as the survey was too long for the given budget.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Benchmarking and accountability during the coronavirus pandemic (#39240)

Created: 04/14/2020 08:30 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
This experiment studies whether and how citizen hold democratic governments accountable during the Covid-19 epidemic. There are two main sets of research questions:

(1) does exogenous variation in information about the response/performance of other countries (what we call benchmarking) affect individuals' beliefs about how well their government has handled the coronavirus? Do benchmarking beliefs have a causal effect on willingness to reward/punish the incumbent government?

Theories of accountability suggest that in order to hold their governments accountable for how they respond to a crisis, voters can rely on (credible) information on how their country performed relative to other countries. Our hypothesis is that exogenous benchmarking information shapes' people's overall evaluation of the government. That is, providing a concrete favorable benchmark positively affects the global evaluation of how well the government has handled the crisis compared to an unfavorable benchmark. A corollary hypothesis is that benchmarking beliefs affect voting behavior.

(2) does an endogenous choice of a benchmark undermine accountability? Is there evidence of political biases in the choice of benchmarks, such that people more (less) inclined to support the government are more (less) likely to select a benchmark favorable to their views? Are people who select a particular benchmark unresponsive to countervailing information?

Theories of political behavior suggest that political pre-dispositions undermine accountability by, among others, affecting information acquisition and/or information processing. In the setting of our experiment with endogenous benchmarking, they predict that pre-treatment political preferences shape benchmark selection.

3) Describe the key dependent variable(s) specifying how they will be measured.
(1a) Assessment of government performance in the crisis: Respondents are asked "Can you tell us how strongly you agree or disagree with the following statement? All in all, the government has handled coronavirus better than most other countries." [Translation from country's language.] Answers are recorded on a 11-point scale (0 = "strongly disagree", 10 = "strongly agree"). Denoted by COMPGOV from now on.

(1b) Vote intentions: Measure 1 (placed several items after experiment in questionnaire) asks respondents how likely it is that their vote is influenced by how the government has handled the coronavirus crisis if an election were held in the near future (next week/Sunday). Responses on 11-point scale (0 = "Very unlikely", 10 = "Very likely"). Measure 2 is a standard vote intention question, which records which party the respondent would vote for if an election were held next week/Sunday. The resulting measure will be equal to 1 if respondents are inclined to vote for the party or parties currently in government, 0 otherwise.

(2) Choice of the benchmark text in treatment condition 3. Binary variable equal to 1 if respondent selects more favorable headline, 0 otherwise.

4) How many and which conditions will participants be assigned to?
Germany, UK, France:

Between-subject design. 3 treatment, 1 control condition.

Control group: receives no benchmarking information
Treatment group 1: receives exogenous benchmarking information indicating that their country's government is doing better in response to the crisis than a benchmark country. Short vignette (no more than 100 words).
Treatment group 2: receives exogenous benchmarking information indicating that their country's government is doing worse in response to the crisis than another country. Short vignette (no more than 100 words).
Treatment group 3: chooses benchmarking information by selecting one of two benchmarking headlines for further reading (positive or negative, as used for treatment groups 1 and 2).

In all treatment conditions respondents are asked to evaluate if text was (i) informative, (ii) credible, and (iii) if the they would share/recommend it.

Austria:

Between-subject design. 1 control condition,1 treatment condition (two stages)

Control group: receives no benchmarking information.
Treatment group: STAGE 1: respondents choose benchmark case by selecting one of two benchmarking headlines for further reading, a positive one (Austria as a leader in fight against coronavirus in Europe) or a negative one (Austria as a laggard). STAGE 2: Among those choosing the positive (negative) benchmarking headline, some receive (weak) counterbalancing information: Austria is a leader in fight against coronavirus in Europe but another country does similarly well (Austria is a laggard but another country in Europe has the same problem).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
(1a) To test the basic benchmarking hypothesis, we regress COMPGOV on treatment indicators, using the negative benchmark as baseline. As stated above, the expectation is that positive benchmarking information leads to an increase in COMPGOV.

(1b) To test the corollary hypothesis regarding vote choice, we will use the fact that the experimental design generates an assignment instrumental variable. Two analyses: (i) An intention-to-treat analysis to estimate the effect of the exogenous benchmark on vote intention (using both measures as dependent variables). Implementation: regress vote intention on treatment indicator variables (with negative benchmark as baseline). (ii) The main quantity of interest is the causal effect of COMPGOV on vote intentions. Implementation: regress vote intentions on COMPGOV instrumented by treatment indicator variables. In IV analysis, we will report results with and without pre-treatment controls for socio-demographics (categories for age, gender, education, current employment status, family structure, region of residence, and current type of housing) as well as pre-treatment measures of news consumption (time spend on political news on an average weekday: none, less than an hour, 1-2 hours, 2-3 hours, more than 3 hours) and trust in media (dummy coding of 4-point scale).

(2) To test the hypothesis concerning the biased choice of benchmark information, we estimate a linear probability model with choice of the favorable benchmark as dependent variable. Beyond socio-demographics and the news consumption measure, an important explanatory variable is the pre-treatment satisfaction with how the executive (prime minister or president) has handled the coronavirus (measured on an 11-point scale).
In the case of Austria, the same analysis of benchmark choice will be conducted. However, given the difference in experimental design the effect of exogenous benchmarking information on the evaluation of government performance will be estimated conditional on choosing a generally positive/negative benchmark.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
No cases will be classified or excluded as "outliers".
In every analysis cases with item non-response will be excluded and reported.
By design, the analysis of endogenous benchmarking can only be conducted for treatment group 3 (treatment group 1 in Austria).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
The experiment is embedded in an opt-in online panel for the cooperative survey project on Citizens' attitudes to Covid-19 run by the international survey company Ipsos. Ipsos will attempt to balance the panel sample to be representative of each country's population of eligible voters.

Target sample sizes:
N=2,000: Germany, France
N=1,000: UK, Austria

Sample size differences are due to resource constraints unrelated to the experiment

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Secondary analyses:

Treatment effect heterogeneity: Does effect of exogenous benchmark treatments on global evaluations vary by trust in media (4-point scale), political news consumption, pre-treatment satisfaction with the prime minister, and pre-treatment satisfaction with how democracy is working in the country?

Related to theories of political behavior, we will assess if respondents exposed to positive (negative) exogenous benchmarking information will be more (less) inclined to evaluate the text positively (informative/credible/willing to share) if they are pre-disposed toward (against) the government.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Does information about comparative vaccination performance matter? (#60659)

Created: 03/11/2021 02:12 PM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
This experiment studies whether and how citizen hold democratic governments accountable during the Covid-19 epidemic with a focus on cross-national benchmarking concerning vaccinations with the possibility of selective exposure. Is there evidence of political biases in the choice of benchmarks, such that people more (less) inclined to support the government are more (less) likely to select a benchmark favorable to their views? What is the effect of exogenous information conditional on prior self-selection into news?

3) Describe the key dependent variable(s) specifying how they will be measured.
1) Choice of a benchmark text among 2 possibilities: for half of the sample, either a neutral or a positive headline; for the other half of the sample, either neutral or negative headline.
(2a) Assessment of government performance in the crisis: Respondents are asked “Can you tell us how strongly you agree or disagree with the following statement? All in all, the government has handled coronavirus better than most other countries.” [Translation from country’s language.] Answers are recorded on a 11-point scale (0 = “strongly disagree”, 10 = “strongly agree”). Denoted by COMPGOV from now on.
(2b) Vote intentions: the variable is a standard vote intention question, which records which party the respondent would vote for if an election were held next week/Sunday. The resulting measure will be equal to 1 if respondents are inclined to vote for the party or parties currently in government, 0 otherwise.
(3) Spending preferences: Respondents are asked “Should there be more or less public expenditure in each of the following areas? Vaccination campaign against COVID19”. Answers are recorded on a 5 point scale : 1. “Much less than now”, 2. “Somewhat less than now” 3. “The same as now” 4. “Somewhat more than now” 5. “Much more than now”.

4) How many and which conditions will participants be assigned to?
Between-subject design. Two stages: headline selection and randomization.
STAGE 1: respondents choose benchmark case by selecting one of two headlines for further reading; Two pairs of headlines are randomly allocated: a neutral one and a positive one (subsample pair1); a neutral one and a negative one (subsample pair2).
STAGE 2: Random allocation of short vignettes with the exact same text (around 1000 characters) and a table comparing France with 4 other OECD countries conditional on selected headline.
Subsample pair1:
-T1. Table with balanced information (France as a middle case among 5 OECD countries).
-T2. Comparatively positive information in the table (France ahead of 5 OECD countries).
Subsample pair2
- T1. Table with balanced information (France as a middle case among 5 OECD countries)
- T3. Comparatively negative information in table (France lagging among 5 OECD countries).

In all treatment conditions respondents are asked to evaluate if text was (i) informative, (ii) credible, and (iii) if the they would share/recommend it.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
(1) To test the hypothesis concerning the biased choice of benchmark information, we estimate regression models with choice of the positive benchmark (relative to neutral) or the negative (relative to neutral) as dependent variables. Test if pre-treatment satisfaction with how the executive has handled the coronavirus is related to selective exposure.

(2) Analysis of benchmarking hypothesis by self-selected strata in stage 1. Depending on subsample, the test concerns the difference between COMPGOV between T2 (T3) and T1 conditional on headline choice. The expectation is that positive (negative) benchmarking information leads to an increase (decrease) in COMPGOV. We also test if there is effect heterogeneity across strata.

(3) To test the corollary hypothesis regarding vote choice and spending preferences on vaccination campaign, we will replicate the analyses described before using vote choice and spending preferences as outcome variables.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

No cases will be classified or excluded as "outliers".

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
The experiment is embedded in an opt-in online panel in France for the project on Citizens' attitudes to Covid-19 run by the international survey company Ipsos. Ipsos will attempt to balance the panel sample to be representative of each country's population of eligible voters.

Target sample sizes:
N=2,000 France

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Treatment effect heterogeneity: Does effect of exogenous benchmark treatments on COMPGOV vary by trust in media (4-point scale), pre-treatment satisfaction with executive, pre-treatment satisfaction with how democracy is working in the country, pre-treatment attitudes towards vaccination?

A.3. Experiment 1

A.3.1. Vignette wording

The list below shows the body of the vignette text presented to respondents. The number of words in each vignette is given in brackets.

France

a. Dans la lutte contre le coronavirus, la France a pris des mesures plus agressives que la Grande-Bretagne. Les deux pays voulaient initialement amortir les coûts économiques du confinement et éventuellement favoriser la création d'une immunité de groupe. Cependant, la France a depuis décidé un confinement très strict tandis que le Président français a souligné que la France a pris "les mesures les plus dures le plus tôt". Alors que dans les deux pays les décès dus à Covid-19 ont augmenté, le Royaume-Uni a connu environ 20 pour cent de décès de plus pour 100 000 habitants. [98]
English translation: In the fight against the coronavirus, France has taken stronger action than Great Britain. The two countries initially wanted to mitigate the economic costs of lockdown and possibly enable the creation of herd immunity. However, France has since decided on a very strict lockdown and the French president said that France took "the toughest measures as soon as possible". While in both countries deaths from Covid-19 have increased, the UK has seen around 20 per cent more deaths per 100,000 population.

b. Dans la lutte contre le coronavirus, la France effectue moins de tests de dépistage que l'Allemagne. L'Organisation mondiale de la santé (OMS) conseille à tous les pays de tester le plus de personnes possible pour dépister le virus. Selon l'OMS, cela permet aux gouvernements de mieux contrôler la propagation du virus et de protéger leurs populations. Le président du Conseil Scientifique a déclaré qu'en France, "nous ne possédons pas les capacités de tester à la même échelle" qu'en Allemagne. Le Gouvernement français a également récemment indiqué que les tests d'anticorps n'étaient pas encore prêts. [96]
English translation: In the fight against the coronavirus, France carries out fewer screening tests than Germany. The World Health Organization (WHO) advises all countries to tests as many people as possible for the virus. According to the WHO, this enables governments to better control the virus and protect their populations. The president of the Scientific Council declared that in France, "we do not have the capacity to test on the same scale" as in Germany. The French government has also recently indicated that antibody tests are not yet ready.

Germany

a. Deutschland führt im Vergleich mit seinen Nachbarn mehr Tests im Kampf gegen das Coronavirus durch. Die Weltgesundheitsorganisation (WHO) rät allen Ländern, möglichst viele Bürger auf den Virus zu untersuchen. Das hilft laut WHO die Epidemie besser zu kontrollieren und die Menschen zu schützen. Deutschland hat im letzten Monat laut aktuellen Schätzungen etwa fünf Mal mehr Tests durchgeführt als Frankreich. [59]
English translation: In the fight against the coronavirus, Germany conducts more tests than its neighbors. The World Health Organization (WHO) advises all countries to tests as many people as possible for the virus. According to the WHO, this enables governments to better control the virus

and protect their populations. Following recent estimates, Germany has conducted approximately five times more tests than France. The Spanish government had to order tests form China to address shortcomings.

b. In Deutschland fehlen im Kampf gegen das Coronavirus Schutzmasken. Die Bundesregierung hat es frühzeitig versäumt, mehr Masken zu besorgen. Gesundheitsminister Jens Spahn hat im einem TV-Interview eingestanden im Februar Hinweise auf mögliche Engpässe nicht weiterverfolgt zu haben. Dagegen hat es Südkorea geschafft, seine Bevölkerung frühzeitig mit Masken zu versorgen. Eine Konsequenz daraus ist, dass eine Lockerung der Kontaktsperre erschwert wird. [60]
English translation: In the fight against the coronavirus, Germany lacks protective face masks. The federal government has failed to acquire more masks early on. Health minister Jens Span admitted in a TV-interview that information about possible shortages was not pursued. In contrast, South Korea has managed to supply its population with face masks. One consequence of the shortage in Germany is that it will be more difficult to relax the lock-down. How and when the lock-down will relaxed in the coming weeks is currently being discussed in Berlin.

United Kingdom

a. In the fight against the coronavirus, the UK has taken more aggressive measures than the Netherlands. Both countries initially took a more conservative approach in order to cushion the economic costs associated with a lockdown and possibly foster the building of herd immunity. However, the UK has since enacted a stricter lockdown. While both countries have seen an increase in deaths from Covid-19, the Netherlands have experienced about 20 percent more deaths per 100,00 inhabitants. [75]

b. In the fight against the coronavirus, the UK conducts less tests than Germany. The World Health Organization (WHO) advises all countries to test as many people as possible for the virus. According to the WHO, this enables governments to better control the virus and protect their populations. The UK government's chief medical officer stated that Germany "got ahead" in testing people. The UK government recently also concluded that some of the antibody tests it ordered abroad were not good to use. [81]

A.3.2. Wording of key survey variables

Below are the question wording and and coding details for the pre-treatment survey questions used in our pre-registered analyses.

Satisfaction with chief executive. This variable is central in our analyses. It measures respondents' (pre-treatment) satisfaction with the head of the executive (we will often refer to this variable with the shorthand "government satisfaction" in the main text). Its wording is as follows: "Generally speaking, are you satisfied or dissatisfied with the action of" {President Macron, Chancellor Merkel, Prime Minister Boris Johnson} Responses are placed on an 11-point scale with labelled endpoints and labelled midpoint ranging from 0 ("completely dissatisfied") to 5 ("neither nor") to 10 ("completely satisfied"). Figure A.1

shows the distribution of this variable in our pooled sample and for each country. While the mean of the satisfaction distribution is rather similar in the pooled sample and in Germany and the UK (around 5.1 in the pooled sample and 5.8 and 5.7 in Germany and the UK, respectively), it is somewhat lower in France (about 4.2). This is because the distribution in France is relatively less left-skewed. When discussing estimates in the main text, we present the marginal effect of a change in satisfaction. However, we also report an alternative quantity that is more sensitive to the underlying satisfaction distribution: the change in the outcome when moving from the 50th percentile of the (country-specific) satisfaction distribution to the 90th percentile. We also conduct (and present in this appendix) semiparametric analyses linking satisfaction to benchmark choice allowing for different satisfaction effect sizes at different levels of satisfaction.

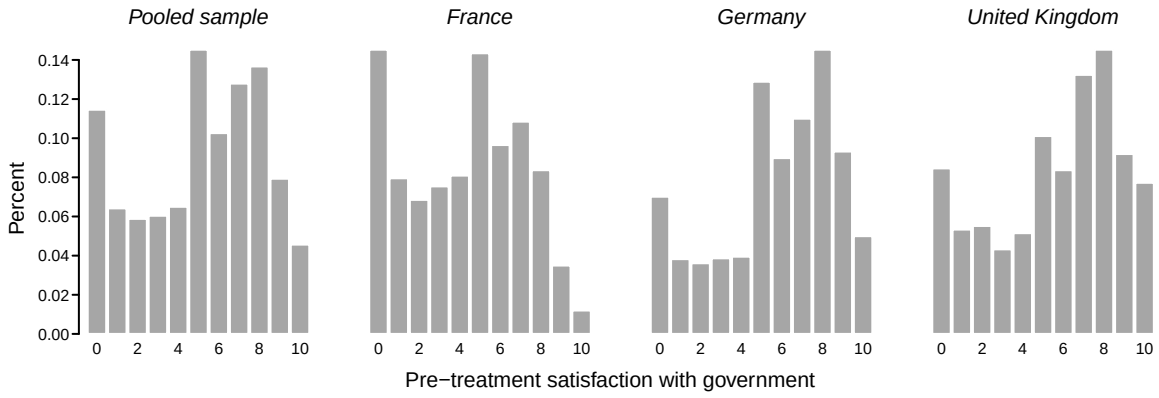

Figure A.1
Histograms of pre-treatment satisfaction with head of the executive

Next, we discuss three variables that were pre-registered for our treatment effect heterogeneity analysis (see section A.3.7).

Trust in the media is measured using question asking respondents to indicate how much they trust journalists on a labelled 4-point scale ranging from "trust completely" to "don't trust at all". "How much do you trust" ... "journalists". Responses are on a labelled 4-point scale comprised of "Trust completely", "trust somewhat", "don't trust a lot', "don't trust at all'. In our analysis we reverse the direction of this variable for ease of presentation.

Political media use is measured using a 4-category item asking respondents how much time they spend on political TV or radio programmes on an average weekday. The exact question wording is: "Roughly speaking, on an average weekday how much time do you spend on": "3. Watching news or political programs on TV" "5. Listen to news or political programs on the radio" Responses are placed in 5 ordered categories: 1. no time, 2. less than 1 hour, 3. 1 to 2 hours, 4. 2 to 3 hours, 5. more than 3 hours. In our analyses

of heterogeneity, we include both ordinal variables in both pseudo-continuous and fully discrete specifications.

Satisfaction with democracy. The exact question wording is: "How satisfied are you with the way democracy works in your country?". Responses are placed on an 11-point scale with labelled endpoints ranging from 0 ("not satisfied at all") to 10 ("very satisfied").

A.3.3. Descriptive statistics of central variables

Table A.2 provides descriptive for experiment 1, across the forced exposure condition (group I) and the choice condition (group II), including pre-treatment satisfaction with the chief executive and the key experimental outcome in each condition. It shows considerable variability in pre-treatment satisfaction, with a standard deviation of 3 around a mean of 5.1 in the pooled sample. Also see Figure A.1.

Table A.2
Descriptive statistics of central variables

	Pooled		France		Germany		UK	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Experiment Group I								
Experimental outcome								
Gov. performance eval.	5.13	2.75	3.81	2.39	6.66	2.38	4.73	2.61
Pre-treatment covariates								
Satisf. w. executive	5.10	3.00	4.21	2.85	5.78	2.84	5.54	3.16
Experiment Group II								
Experimental outcome								
Pos. headline choice	0.31	0.46	0.31	0.46	0.33	0.47	0.29	0.45
Pre-treatment covariates								
Satisf. w. executive	5.16	2.99	4.31	2.80	5.77	2.95	5.67	3.05

A.3.4. Respondent evaluations of experiment

Panel (A) of Figure A.2 shows respondents' mean rating of how informative and credible they perceive a vignette to be in each country (averaging over all experimental groups). Panel (B) shows respondents' mean rating of how informative and credible they perceive a vignette to be separately for experimental conditions *Ia*, *Ib*, and *II*. Panel (C) shows mean ratings of respondents in experimental group *II* only, separated by their choice of positive or negative benchmark headlines.

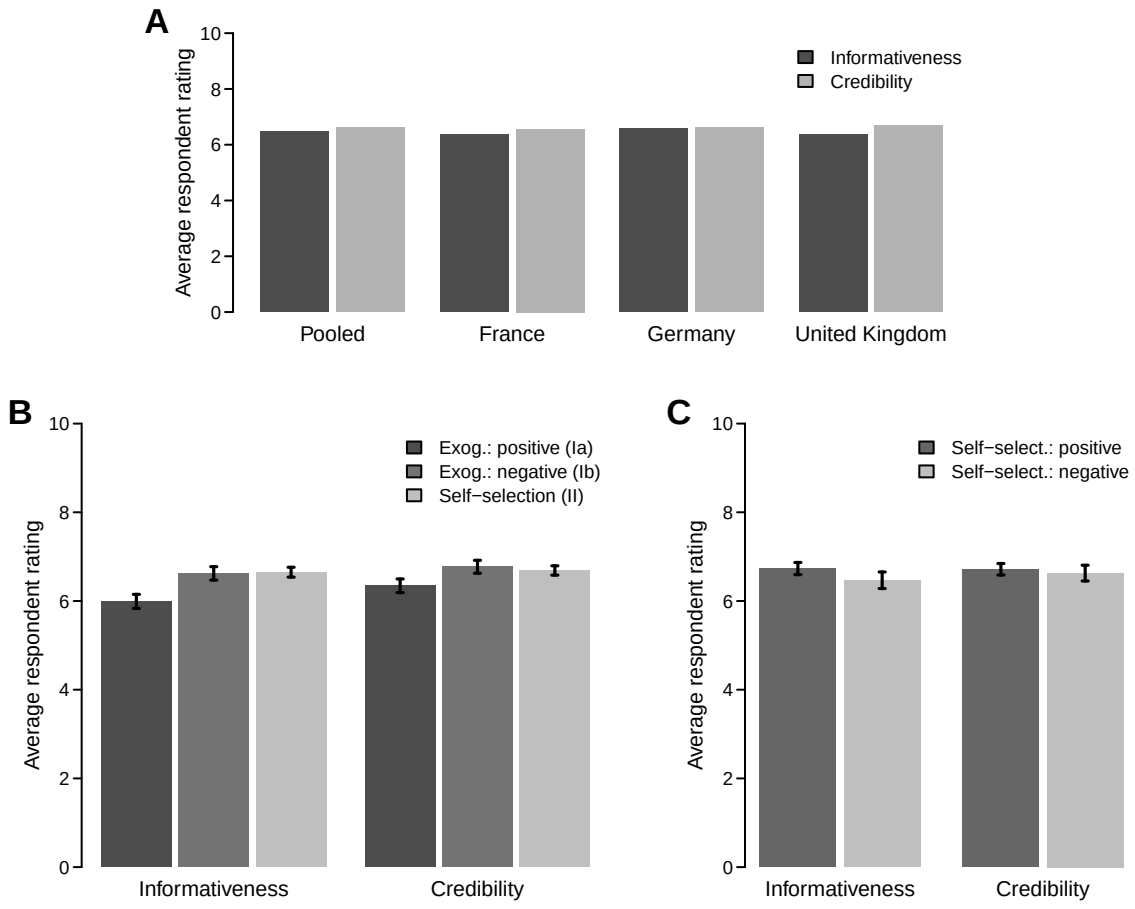

Figure A.2
Respondent evaluations of vignettes.

Barplots of average respondent ratings of informativeness and credibility of vignettes. Panel (A) plots ratings by country averaging over all experimental groups. Panel (B) compares ratings among the three experimental groups. Panel (C) compares ratings by choice of benchmark headline in group II. Means weighted by sample inclusion probability. Error bars show 95% confidence intervals.

A.3.5. Additional analysis of endogenous benchmark choice

In this section, we present results from a series of models that semiparametrically estimate the relationship between pre-treatment government satisfaction and the choice of a positive benchmark headline. To do so, we estimate generalized additive logit models (Hastie and Tibshirani 1986; Beck and Jackman 1998), where the effect of satisfaction is modeled via thin-plate regression splines (Wood 2003). Figure A.3 plots conditional predicted probabilities (on the y-axis) against the range of satisfaction (on the x-axis). It reveals that the effect of satisfaction on benchmark choice is fairly linear across the range of satisfaction, especially in Germany and the UK, so that the marginal effects reported in the main text are a sensible one-number-summary measure.

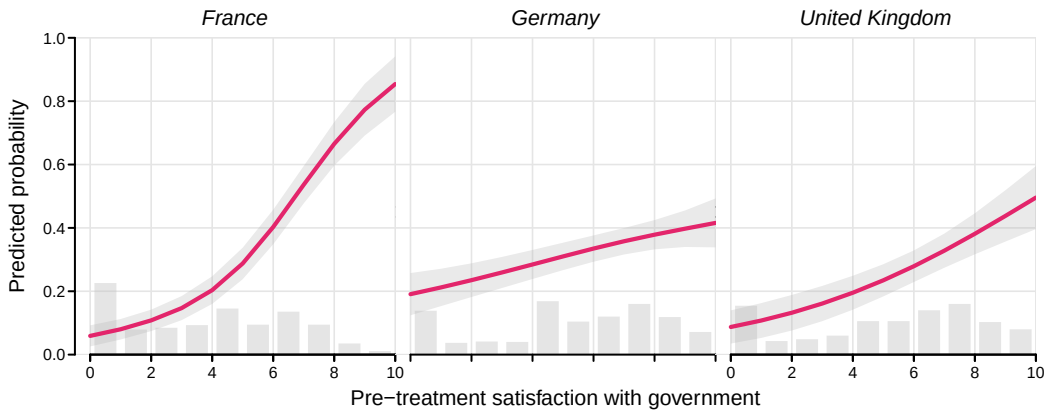

Figure A.3
Semiparametric model of the probability of positive benchmark choice as function of pre-treatment satisfaction with government

Shown are predicted probabilities (with 95% confidence intervals) calculated from generalized additive logit models with non-linear terms for government satisfaction effect estimated via penalized thin-plate regression splines. The distribution of satisfaction is shown as grey histogram bars above the x-axis.

A.3.6. Estimates of exogenous benchmark effect

Table A.3 shows estimates of the average treatment effect of exogenous benchmark provision on respondents' government performance evaluations expressed in various units. First, we display the ATE on the original scale of the survey variable (ranging from 0 to 10), Next we display the ATE expressed in standard deviation units. We also express the magnitude of the ATE as a percentage increase of the respective sample mean. The final reported quantities are p-values testing the sharp null hypothesis of no treatment effect using randomization test. Panel (A) of Table A.3 shows results without covariates, while panel (B) shows results when adjusting for pre-treatment survey design variables.

Table A.3
Effect of exogenous benchmark on government performance evaluation

	Pooled	Germany	France	United Kingdom
A: Average treatment effects				
ATE [on 0-10 scale]	0.300	0.272	0.300	0.367
	(0.125)	(0.173)	(0.177)	(0.263)
ATE [in SD units]	0.109	0.114	0.125	0.141
	(0.046)	(0.072)	(0.074)	(0.101)
ATE [change in %]	6.01	4.17	8.12	8.08
Randomization p-value	0.003	0.041	0.078	0.081
B: Covariate-adjusted average treatment effects				
ATE [on 0-10 scale]	0.305	0.251	0.299	0.346
	(0.125)	(0.169)	(0.176)	(0.260)
ATE [in SD units]	0.111	0.105	0.125	0.132
	(0.045)	(0.071)	(0.073)	(0.099)
ATE [change in %]	6.12	3.85	8.07	7.60
Randomization p-value	0.002	0.055	0.085	0.096

Note: This table shows the average treatment effect of exogenous benchmark provision on performance evaluations. It provides estimates expressed in several different units: on the original scale (0-10) of the survey variable, in standard deviation units, and as percentage change from the sample mean. Panel (A) shows results without covariates, while panel (B) shows results when adjusting for pre-treatment survey design variables (age, gender, education, and employment status). Robust standard errors in parentheses. Randomization p values are based on 1,000 draws.

A.3.7. Treatment effect heterogeneity

In this section, we present analyses testing for heterogeneous treatment effects. We test for heterogeneity in treatment effects due to pre-treatment measures of satisfaction with the chief executive, satisfaction with democracy, media usage and trust in the media (based on pre-registered hypotheses). In Table A.4, panel (A) we report randomization p-values testing the sharp null hypothesis of a constant treatment effect using linear interactions of the treatment variable with the pre-treatment covariates. To guard against the linear functional form assumptions driving these findings (Hainmueller, Mummolo and Xu 2019) we also present nonlinear interactions in panel (B), where we interact the treatment with each observed category to create a completely non-linear interaction surface. Because we are carrying out a multitude of significance tests, it is prudent to adjust p-values for multiple testing in order to guard against false positive findings. In the rightmost set of columns of Table A.4, we thus report randomization p-values adjusted for multiple testing so that the family-wise error rate (the probability of at least one false positive among the set of tests) is at most 5% using the Holm-Bonferroni (Holm 1979) methodology.

Table A.4
Examining treatment effect heterogeneity in pre-treatment covariates. Randomization tests, p-values (without and with adjustment for multiple-testing)

	p-values				p-values, *FWER*-adjusted			
	All	FR	DE	UK	All	FR	DE	UK
A: Linear interaction models								
Satisfaction with gov.	0.66	0.76	0.46	0.27	1.00	1.00	1.00	1.00
Political media usage	0.49	0.65	0.56	0.40	1.00	1.00	1.00	1.00
Trust in the media	0.16	0.48	0.85	0.22	0.65	0.96	0.96	0.65
Satisfaction with Dem.	0.59	0.60	0.84	0.80	1.00	1.00	1.00	1.00
B: Non-linear interaction models								
Satisfaction with gov.	0.89	0.38	0.88	0.47	1.00	1.00	1.00	1.00
Political media usage	0.94	0.68	0.71	0.88	1.00	1.00	1.00	1.00
Trust in the media	0.32	0.57	0.05	0.25	0.74	0.74	0.21	0.74
Satisfaction with Dem.	0.11	0.61	0.91	0.16	0.42	1.00	1.00	0.48

Note: Based on 10,000 randomized treatment assignments in treatment-by-covariate interaction models testing the sharp null hypothesis of a constant average treatment effect. Pooled sample results calculated assuming randomization blocked by country. Panel A shows the resulting p-values when using linear interaction terms, panel B shows p-values of models allowing for non-linearity in the interaction surface, where we interact the treatment with each observed value of the variable. *FWER*-adjusted p values are adjusted for multiple testing to have a family wise error rate of at most 5% using the Holm-Bonferroni method (Holm 1979).

We do not find evidence for heterogeneous treatment effects. Faced with the same benchmarked news on the pandemic, respondents with different prior political beliefs, media usage, trust in the media, or satisfaction with democracy did not tend to evaluate government performance in a significantly different way. In other words, we cannot reject the null hypothesis of a constant treatment effect for any the four variables considered. Note that trust in the media in Germany using a categorical interaction produces a p-value of 0.05. However, when adjusting for multiple testing, the corresponding p-values is 0.21. We thus think it prudent to conclude that no clear evidence for effect heterogeneity is found in our sample.[1]

Table A.5
Additional analyses of treatment effect heterogeneity. Respondents' views on consequences of Coronavirus for health and economy.

	p-values				p-values, *FWER*-adjusted			
	All	FR	DE	UK	All	FR	DE	UK
A: Linear interaction models								
Coronavirus: health	0.97	0.66	1.00	0.45	1.00	1.00	1.00	1.00
Coronavirus: economy	0.68	0.93	0.49	0.31	1.00	1.00	1.00	1.00
B: Non-linear interaction models								
Coronavirus: health	0.14	0.06	0.12	0.36	0.35	0.26	0.35	0.36
Coronavirus: economy	0.35	0.30	0.39	0.27	1.00	1.00	1.00	1.00

Note: Randomization tests, p-values (without and with adjustment for multiple-testing). For construction details see Table A.4.

Table A.5 explores an additional dimension of heterogeneity: respondents' assessment of the severity of the impact of the pandemic. Note that we did not pre-register these analyses. Instead they arose during the review process, and we report them here due to their substantive importance. Individual differences in beliefs about the likely impact of the crisis on public health and the economy might moderate the impact of our experimental treatment effect. We thus conducted further tests of treatment effect heterogeneity using two survey items with which we probed how serious respondents thought the consequence of the Coronavirus pandemic were for health and the economy of their country.[2] There is

[1] The sample sizes for these analyses are about 800 in France and Germany, and about 400 in the UK. Thus, it is of course possible that effect heterogeneity can be detected in future studies employing much larger samples sizes.

[2] The exact question wording is: "Would you say the consequence of the Coronavirus epidemic for health in country / for country's economy are..."Response options were (1) Very serious, (2) Quite serious, (3) Somewhat serious, (4) Not serious, (5) Not at all serious. Very few respondents viewed the consequences as "not at all serious", thus, we collapsed response categories 4 and 5.

substantial variation among individuals. In Germany, about 36 percent of respondents think that the consequences of the pandemic are only "somewhat serious" or not at all serious for public health. The corresponding percentages are 14 and 11 percent in France and the UK.[3] However, as the reported randomization p-values in Table A.5 show, we find no clear evidence that the benchmarking treatment effect is heterogenous in existing beliefs about the impact of the pandemic.

A.3.8. Impact of country references in vignette headlines

As discussed in the main text (recall Table 1), in experiment 1 the headline in Germany differs from the two other countries in that it does not mention a reference country. Table A.6 reports a test whether the effect of exogenous information varies between vignette headlines with and without country labels. We find no evidence of such heterogeneity.

Table A.6
Randomization test of ATE heterogeneity contrasting vignette headlines with and without country labels.

	F	p
H_0: constant ATE	0.007	0.924

Note: Randomization inference based on 10,000 block (by country) randomized treatment schedules. F-test of difference of average treatment effect contrasting Germany (no country names in headlines) to France and the UK.

A.3.9. Benchmarks, performance evaluations, and vote choice

The analysis in the main text focuses, first, on the effect of exogenous benchmarking on performance evaluation capturing whether the government has handled the pandemic well compared to most other countries, and, second, on the relevance of endogenous benchmarking through self-selection into benchmarking headlines. What about the link between benchmarking, performance evaluations, and the vote? The analyzes summarized in Figure A.4 address this question.

Panel **a** shows that in all three countries under study individuals who think that the government has handled the crisis comparatively well are more much more likely to indicate that they would vote for the government if parliamentary elections happened next Sunday compared to those who think the government has not handled the crisis well.

[3] Respondents are somewhat less sanguine about economic consequences: the percentages are 15 in Germany, and 8 and 7 in France and the UK.

While vote intention is measured well after the experiment at the end of the survey, this does not rule out reverse causality or omitted confounders (such as partisanship).

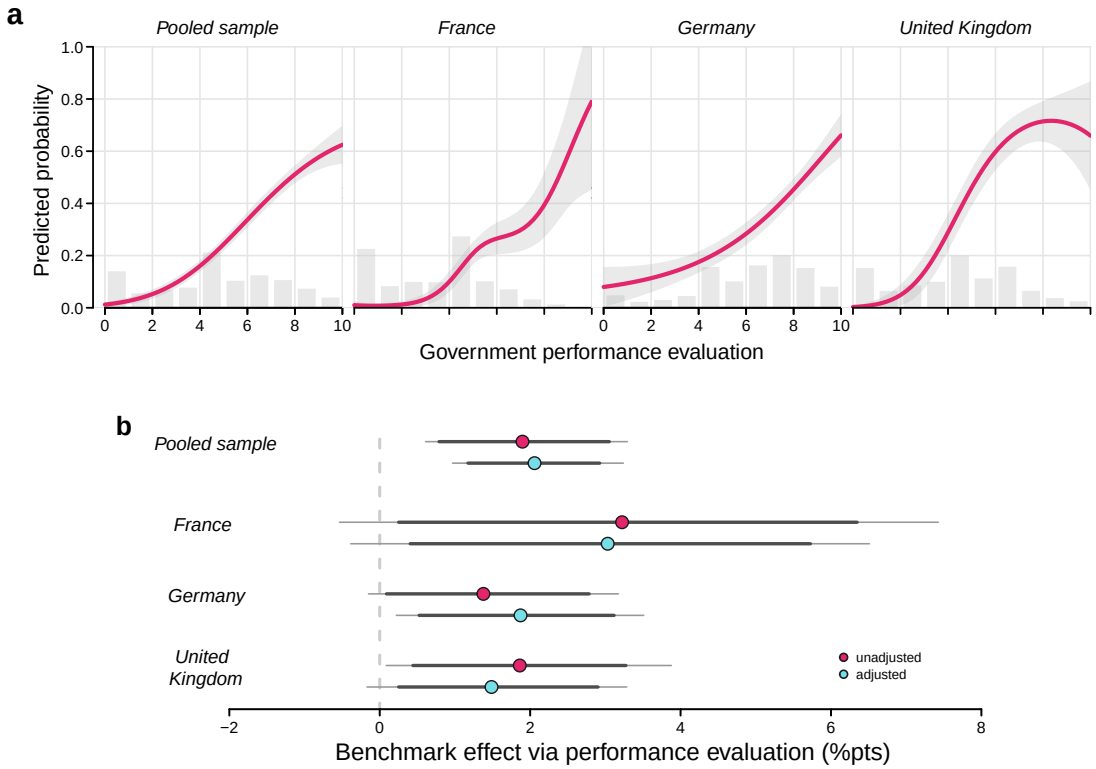

Figure A.4
Benchmarking information, performance evaluations, and vote choice

Panel (**a**) plots the relationship between government performance evaluations and the stated intention to vote for the governing party or coalition in the next election. Predicted probabilities (with 95% confidence intervals) calculated from generalized additive logit models with non-linear terms for government performance evaluation estimated via penalized thin-plate regression splines. The distribution of performance evaluation is shown as grey histogram bars above the x-axis. Panel (**b**) plots the impact of an exogenous change in positive benchmarking information on vote intention channeled ('mediated') via changes in performance evaluation. Plotted are differences in predicted probabilities of vote intention (in %pts) without covariate adjustment (●) and with covariate adjustment (○). Confidence intervals (with 90% and 95% coverage) based on nonparametric bootstrap (500 draws). Mediated effect estimates calculated following Imai, Keele and Tingley (2010). The outcome equation uses the same generalized additive model as in (a) with an additional coefficient for the randomized treatment. The mediator equation is a linear model regressing performance evaluations on randomized benchmark treatment.

The analysis in Panel **b** of Figure A.4 more formally investigates the theoretical channel from benchmarking information to vote choice. Using only respondents in the exogenous information condition, we estimate the effect of the positive compared to the negative benchmarking information on vote intention channeled ('mediated') through the overall evaluation of government performance in the pandemic. This is what the literature usually

calls the natural indirect effect or average causal mediation effect. Our estimation method follows the procedure proposed by Imai, Keele and Tingley (2010). In the pooled model, the estimates suggest that positive benchmarking information increases the probability of voting for the government through changing performance evaluations by 1.8 percentage points. The confidence intervals are sufficiently narrow to conclude that this mediation effect is statistically significantly different from zero. The result is essentially the same with and without covariates. Covariates include age, gender, university education, employment status, trust in the media, and political media usage (the maximum of consumption of political programs on TV or radio), and region of residence. This causal mediation analysis does not require an exclusion restriction, that is, there may be direct effect of the treatments on vote choice via other channels. However, a causal interpretation of the mediation effect is not justified by the experiment alone. Randomization the treatment only ensures the exogeneity of the treatments, but does not address omitted variables shaping both the mediator, performance evaluations, and the outcome variable. However, it is reassuring that adjusting for possible confounders does not substantively change the estimated mediation effect.

A.3.10. Additional experiment: Austria

Austria implemented a different version of experiment 1. There is no purely exogenous information condition. First, all respondents participating in the experiment are asked to choose one of the (benchmarking) headlines for further reading, a positive one (Austria as a leader in fight against coronavirus in Europe) or a negative one (Austria is a laggard in providing tests). This enables us to test for endogenous benchmarking.

Second, conditional on the benchmarking choice, we randomize whether respondents receive (weak) counterbalancing information. This conditional randomization enables us to test for the impact of countervailing information conditional on self-selection. The full text of the vignettes (in German) is provided below. Respondents who selected the positive headline always got a positive vignette text in line with the headline, including comparative information on lockdown-style measures and praise by German chancellor Angela Merkel. But some vignettes note that another country (i.e., South Korea) does similarly well. The idea is to provide information that may lead to a marginal adjustment in relative performance evaluations conditional on positive selection. Respondents who selected the negative headline got a vignette text elaborating on the headline. It notes that Austria lags behind in testing compared to Germany, which has conducted about three times the number of tests per 100,000 inhabitants. But some vignettes note that another country in Europe (i.e., France) has the same problem.

Vignette wording in Austria After selecting a headline, respondents are asked to read the corresponding text and answer questions. Each respondent only sees one vignette.

Choice of negative headline: Österreich hinkt beim Testen hinterher (Austria lags behind in testing)

a. Im Kampf gegen das Coronavirus hinkt Österreich beim Testen Deutschland hinterher. Die Weltgesundheitsorganisation (WHO) rät allen Ländern, möglichst viele Bürger auf den Virus zu untersuchen. Das hilft laut WHO die Epidemie besser zu kontrollieren und die Menschen zu schützen. Bundeskanzler Sebastian Kurz proklamierte zwar: "Testen, testen, testen." Doch Deutschland hat im letzten Monat laut aktuellen Schätzungen etwa drei Mal mehr Tests pro 100.000 Einwohner durchgeführt als Österreich. Auch Südkorea hat frühzeitig und umfangreich getestet und steht besser da als viele andere Länder.

English translation: In the fight against the coronavirus, Austria is lagging behind Germany in testing. The World Health Organization (WHO) advises all countries to tests as many people as possible for the virus. According to the WHO, this enables governments to better control the virus and protect their populations. Chancellor Sebastian Kurz proclaimed: "Test, test, test." But according to current estimates Germany carried out about three times more tests per 100,000 inhabitants than Austria in the last month. South Korea also tested early and extensively and is in a better position than many other countries.

b. Im Kampf gegen das Coronavirus hinkt Österreich beim Testen Deutschland hinterher. Die Weltgesundheitsorganisation (WHO) rät allen Ländern, möglichst viele Bürger auf den Virus zu untersuchen. Das hilft laut WHO die Epidemie besser zu kontrollieren und die Menschen zu schützen. Bundeskanzler Sebastian Kurz proklamierte zwar: "Testen, testen, testen." Doch Deutschland hat im letzten Monat laut aktuellen Schätzungen etwa drei Mal mehr Tests pro 100.000 Einwohner durchgeführt als Österreich. In anderen europäischen Ländern, wie beispielsweise in Frankreich, gibt es auch Engpässe bei Tests.
English translation: In the fight against the coronavirus, Austria is lagging behind Germany in testing. The World Health Organization (WHO) advises all countries to tests as many people as possible for the virus. According to the WHO, this enables governments to better control the virus and protect their populations. Chancellor Sebastian Kurz proclaimed: "Test, test, test." But according to current estimates Germany carried out about three times more tests per 100,000 inhabitants than Austria in the last month. In other European countries, such as France, there are also bottlenecks in testing.

Choice of positive headline: Österreich ist Taktgeber Europas (Austria is Europe's pace setter

a. Im Kampf gegen das Coronavirus hat Österreich schneller auf einen nationalen Shutdown gesetzt als Deutschland. Laut einer Analyse der Universität von Oxford hat Österreich bis Ende März einen umfangreicheren Maßnahmenkatalog zur Eindämmung des Virus umgesetzt. Dieser beinhaltet mehr Einschränkungen für den Alltag der Menschen. Der Erfolg der Maßnahmen erlaube es laut der Bundesregierung in Wien, das öffentliche Leben jetzt schrittweise wieder hochzufahren. Auch mit der angekündigten Lockerung des Shutdowns ist Österreich Taktgeber in Europa. „Österreich war uns immer einen Schritt voraus," so die deutsche Bundeskanzlerin.
English translation: In the fight against the coronavirus, Austria was more rapid than Germany in enacting a national lockdown. According to an analysis by the University of Oxford, by the end of march Austria had implemented a more extensive catalogue of measures to contain the virus. It includes more restrictions on people's everyday lives. The success of the measures now makes it possible to gradually start up public life again, according to the federal government in Vienna. Also with the announced easing of the lockdown, Austria is Europe's pacesetter. "Austria was always one step ahead of us," said the German Chancellor.

b. Im Kampf gegen das Coronavirus hat Österreich schneller auf einen nationalen Shutdown gesetzt als Deutschland. Laut einer Analyse der Universität von Oxford hat Österreich bis Ende März einen umfangreicheren Maßnahmenkatalog zur Eindämmung des Virus umgesetzt. Dieser beinhaltet mehr Einschränkungen für den Alltag der Menschen. Der Erfolg der Maßnahmen erlaube es laut der Bundesregierung in Wien, das öffentliche Leben jetzt schrittweise wieder hochzufahren. Auch mit der angekündigten Lockerung des Shutdowns ist Österreich Taktgeber in Europa. In Asien hat Südkorea frühzeitig reagiert und steht besser da als viele andere Länder.
English translation: In the fight against the coronavirus, Austria was more rapid than Germany in enacting a national lockdown. According to an analysis by the University of Oxford, by the end of march Austria had implemented a more extensive catalogue of measures to contain the virus. It includes more restrictions on people's everyday lives. The success of the measures now makes it possible to gradually start up public life again, according to the federal government in Vienna. Also

with the announced easing of the lockdown, Austria is Europe's pacesetter. In Asia, South Korea reacted early and is doing better than many other countries.

Results Figure A.5 presents the results. Panel (a) replicates the analysis of benchmark choice. As in France, Germany, and the UK, we find that pre-treatment satisfaction with the chief executive is a significant predictor of benchmark choice. People who were more satisfied with chancellor Sebastian Kurz before seeing and choosing headlines were more likely to pick the headline indicating a positive benchmark. The slope of the relationship is steeper than in the Germany or the UK and similar to France. Altogether, we find clear evidence of endogenous benchmarking based on political characteristics in each of the four countries we study, covering different types of parliamentary regimes, some more majoritarian and other more consensual, and with varying degree of party polarization.

Panel (b) of Figure A.5 plots the effect of providing counterbalancing information after benchmark choice. Given the selection of a negative headline, respondents who received the corresponding text describing Austria as a laggard in testing but with some counterbalancing information have, on average, a marginally higher evaluation of the government's comparative performance than respondents that do not receive any counterbalancing information. However, as the length of the standard error bar indicates, this difference-in-means of 0.26 (4.3% of the mean in the other group) is clearly not statistically significant. The difference is a little bit smaller than the effects of unconditional exogenous information found in the main version of experiment 1. Conditional on the selection of a positive headline, there is no difference in the performance evaluations based on whether respondents receive some counterbalancing information.

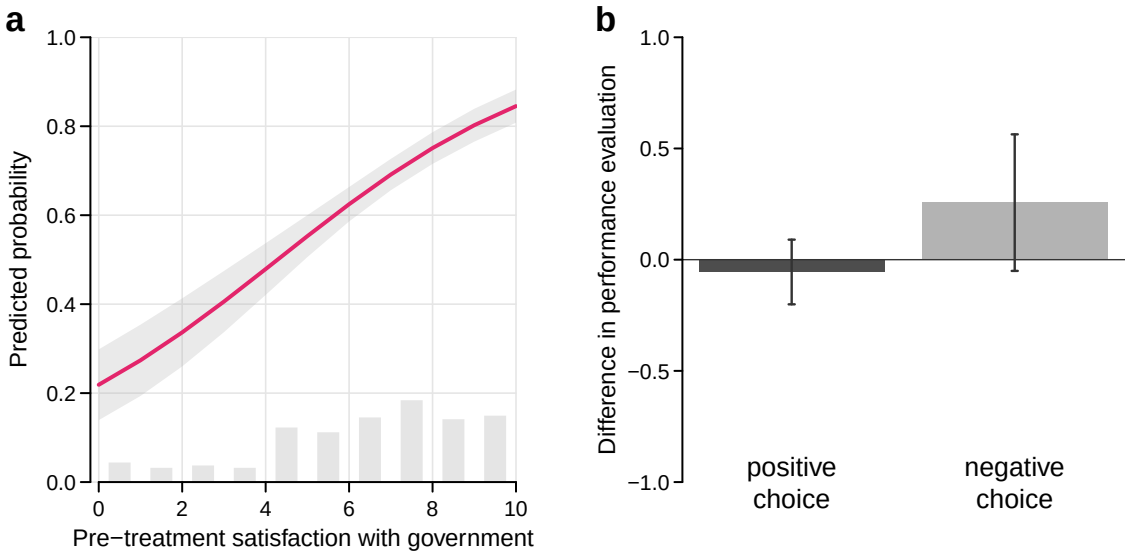

Figure A.5
Benchmark choice and information treatment effects in Austria

Panel (**a**) plots the probability of a respondent choosing a positive benchmark headline as a function of pre-treatment government satisfaction. Predicted probabilities (with 95% confidence intervals) calculated from generalized additive logit models with non-linear terms for government satisfaction effect estimated via penalized thin-plate regression splines. The distribution of satisfaction is shown as grey histogram bars above the x-axis. Panel (**b**) plots the effect of providing counterbalancing information after benchmark choice. Bars are treatment-control group differences (weighted by sample inclusion probability), error bars show robust standard errors.

A.4. Experiment 2

A.4.1. Vignette wording

All three vignettes have the same introductory text:

> Alors que de nombreux pays ont débuté leur campagne de vaccination contre le coronavirus fin 2020, comment se situe comparativement la proportion de personnes vaccinées en France?
>
> Le coronavirus fait toujours rage dans le monde! Le nombre de cas quotidien ne cesse de battre des records et de nouveaux variants sont détectés aux quatre coins de la terre. Alors que certains pays se désespèrent, d'autres ont pu débuter leur campagne de vaccination depuis le mois de décembre 2020. Depuis le début de la pandémie, les experts parlent d'une possible immunité collective une fois que 60% de la population sera immunisée.
>
> Quel pourcentage de la population est déjà vacciné dans 5 pays de l'OCDE ayant débuté la vaccination ? Le calcul est basé sur le nombre de personnes ayant reçu au moins une première dose de vaccin dans chaque pays.

English translation:

> Now that many countries have started their vaccination campaign against the coronavirus at the end of 2020, how does the proportion of people vaccinated in France look in a comparative perspective?
>
> The coronavirus is still raging around the world! The number of daily cases continues to break records and new variants are detected in the four corners of the earth. While some countries are in despair, others have been able to start their vaccination campaign since December 2020. Since the start of the pandemic, the experts speak of a possible collective immunity once 60% of the population is immunized.
>
> What percentage of the population is already vaccinated in 5 OECD countries that have started vaccination? The calculation is based on the number of people who received at least a first dose of vaccine in each country.

Table A.7 shows the benchmarking information tables presented to respondents (depending on random assignment in stages I and III of the experiment). Each table shows vaccination rates for five OECD countries, including the respondent's home country (France) at the time of the survey. In the positive benchmarking information treatment, France is compared favorably to four vaccination laggards. In the negative case, France is placed last compared to four vaccination leaders. In the neutral case, France is compared to one leader, one laggard and two neighboring countries with similar vaccination rates.

Table A.7
Benchmarking information used in experiment 2.

IIIa. Positive benchmarking information

Pays	Personnes vaccinées	Population totale	Pourcentage de personnes vaccinées
France	**3.9 millions**	**67 millions**	**5.8%**
Canada	1.8 millions	37.6 millions	4.9%
Autriche	0.3 millions	8.9 millions	3.8%
Corée du Sud	0.3 millions	51.7 millions	0.6%
Australie	0.01 millions	25.3 millions	0.3%

IIIb. Neutral benchmarking information

Pays	Personnes vaccinées	Population totale	Pourcentage de personnes vaccinées
Royaume-Uni	22.4 millions	66.6 millions	33.6%
Allemagne	5.2 millions	83 millions	6.2%
France	**3.9 millions**	**67.0 millions**	**5.8%**
Belgique	0.6 millions	11.5 millions	5.4%
Australie	0.01 millions	25.3 millions	0.3%

IIIc. Negative benchmarking information

Pays	Personnes vaccinées	Population totale	Pourcentage de personnes vaccinées
Royaume-Uni	22.4 millions	66.6 millions	33.6%
États-Unis	60 millions	382.2 millions	18.3%
Danemark	0.5 millions	5.8 millions	9.1%
Espagne	3.3 millions	46.9 millions	7.1%
France	**3.9 millions**	**67 millions**	**5.8%**

Note: Decimal commas have been converted to decimal points for consistency of presentation.

A.4.2. Additional results

Table A.8 shows the proportion of respondents that chose a directional (positive or negative) headline in the second experiment. The last column shows exact p-values from binomial proportion tests of the null hypothesis that respondents select headlines at random. It is noteworthy that respondents are clearly less likely to select positive headlines. Only about one third of respondents chose a positive over a neutral headline

in group *Ia*, which is rather close to the proportion found in the first experiment (0.31), which contrasted positive to negative headlines.

Table A.8
Test of non-random benchmark choice in experiment 2.

	Choice proportion	$H_0 : Pr = 0.5$
Ia: positive vs. neutral headline	0.320 (*IIa*)	0.000
Ib: negative vs. neutral headline	0.445 (*IIb*)	0.001

Table A.9 shows group means and differences for the second experiment. Panel (A) shows raw experimental group means and differences, while panel (B) adjusts for individual pre-treatment covariates. Like in our other analyses, we include a respondent's gender, age, education (having completed a BA or above), and employment status. Our conclusions are not altered by covariate adjustment.

A.4.3. Additional analysis of endogenous benchmark choice

The marginal effects presented in Figure IV in the main text are based on a linear probability model and assume constant marginal effects of satisfaction. To allow for a more flexible assessment of the relationship between pre-treatment satisfaction and headline choice, we also estimate a set of semi-parametric models. Figure A.6 plots predicted probabilities of respondents choosing a positive/negative headline in stage II of the second experiment. We find that respondents who are more satisfied with the government to begin with are more likely to choose a positive headline. The estimates imply that strong supporters of the government are about three-times as likely to choose a positive over a neutral headline than strong opponents of the government. The quantitative magnitude is somewhat smaller in the second experiment compared to the first experiment for France, likely representing the weaker contrast of the choice options (positive-neutral versus positive-negative). Though the gap remains substantively large. We find commensurate evidence of non-random selection of negative headlines. As one would expect, respondents that are more satisfied with the performance of the executive are less likely to select negative headlines.

Table A.9
Benchmark choice, exogenous benchmarking information, and evaluation of government performance.

A: Unadjusted means			
	Neutral versus positive headline condition		
	neutral choice	positive choice	Difference
Balanced information	3.64	4.34	−0.70 (0.25)
Positive information	3.79	4.46	−0.67 (0.24)
Difference	0.15 (0.19)	0.12 (0.29)	0.03 (0.35)
	Neutral versus negative headline condition		
	neutral choice	negative choice	Difference
Balanced information	4.63	2.94	1.69 (0.21)
Negative information	4.27	2.59	1.68 (0.22)
Difference	−0.36 (0.20)	−0.35 (0.23)	−0.01 (0.30)
B: Adjusted for covariates			
	Neutral versus positive headline condition		
	neutral choice	positive choice	Difference
Balanced information	3.63	4.34	−0.71 (0.25)
Positive information	3.78	4.52	−0.74 (0.24)
Difference	0.15 (0.19)	0.18 (0.29)	−0.03 (0.34)
	Neutral versus negative headline condition		
	neutral choice	negative choice	Difference
Balanced information	4.63	2.94	1.69 (0.21)
Negative information	4.27	2.59	1.68 (0.22)
Difference	−0.36 (0.20)	−0.35 (0.23)	−0.01 (0.30)

Note: Panel (A) shows raw experimental group means and differences. Panel (B) shows adjusted means and differences, adjusting for individual differences in age, gender, education, and employment status. Weighted by sample inclusion probability. Robust standard errors in parentheses.

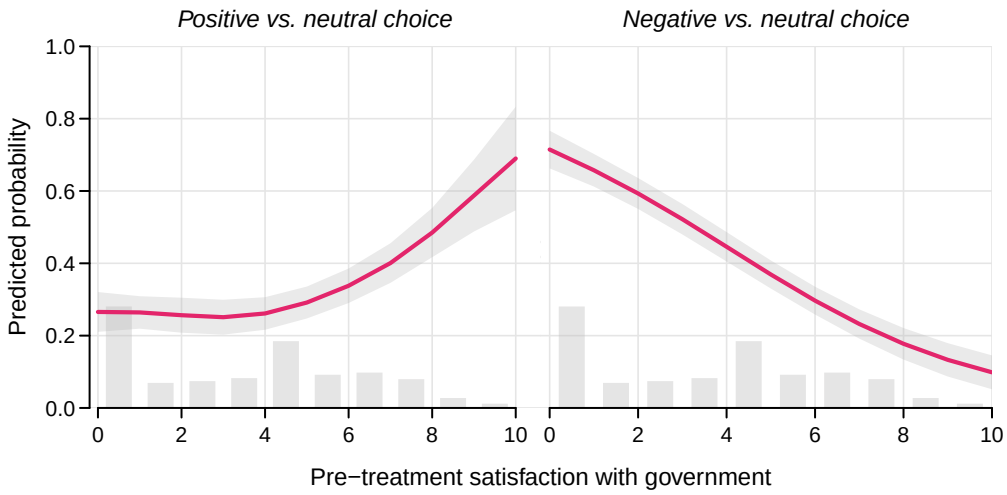

Figure A.6

Pre-treatment government satisfaction and benchmark headline selection in experiment 2.

This figure plots the probability (with 95% confidence intervals) of a respondent choosing a positive (left panel) or negative (right panel) benchmark headline over the neutral alternative as a function of pre-treatment government satisfaction. Experiment 2 conducted in France. Predicted probabilities calculated from generalized additive logit models with non-linear terms for government satisfaction effect estimated via penalized thin-plate regression splines. The distribution of satisfaction is shown as grey histogram bars above the x-axis.

References

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27:163–192.

Hastie, Trevor and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1(3):297–310.

Holm, Sture. 1979. "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* 6(2):65–70.

Imai, Kosuke, Luke Keele and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological Methods* 15(4):309–334.

Wood, Simon N. 2003. "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B* 65(1):95–114.