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

Michael Becher[a] Sylvain Brouard[b] Daniel Stegmueller[c]

March 27, 2023
Accepted for publication at
British Journal of Political Science

Abstract

When do cross-national comparisons enable citizens to hold governments accountable? According to recent work in comparative politics, benchmarking across borders is a powerful mechanism for making elections work. However, little attention has been paid to the choice of benchmarks and how it shapes democratic accountability. We extend existing theories to account for *endogenous* benchmarking. Using the COVID-19 pandemic as a test case, we embedded experiments capturing self-selection and exogenous exposure to benchmarked information in representative surveys in France, Germany, and the UK. The experiments reveal that when individuals have the choice, they are likely to seek out congruent information in line with their prior view of the government. Going beyond existing experiments on motivated reasoning and biased information choice, endogenous benchmarking occurs in all three countries despite the absence of partisan labels. Altogether, our results suggest that endogenous benchmarking weakens the democratic benefits of comparisons across borders.

Key words: information choice; benchmarking; blame attribution; accountability; motivated reasoning; COVID-19

[a]IE University, Madrid, Spain; Institute for Advanced Study in Toulouse, Toulouse, France. Corresponding author. Email: michael.becher@ie.edu
[b]Sciences Po, Center for socio-political data (CDSP) & Center for Political Research (CEVIPOF), CNRS, Paris, France. Email: sylvain.brouard@sciencespo.fr
[c]Duke University, Durham, NC, USA. Email: daniel.stegmueller@duke.edu

A vast literature in political science remains divided over whether retrospective evaluations of government performance by citizens can provide a reliable basis for substantive electoral accountability. While free and fair elections constitute a formal link of accountability between citizens and elected policymakers, substantive accountability means that elections are an instrument of selecting competent policymakers and/or incentivizing incumbents to exert effort in the public interest. An important part of the debate focuses on how individuals use (or fail to use) the information required to appropriately assign responsibility for government performance.[1]

While evaluating government performance is a complex task, benchmarking theories of accountability argue that cross-national comparisons provide a useful and easily available heuristic for citizens (Kayser and Peress 2012; Park 2019; Powell and Whitten 1993). In particular, benchmarked information in the media can provide the needed input for the process of democratic accountability. For example, if citizens learn that their country has provided more coronavirus tests or vaccinations than a comparison country during the COVID-19 pandemic, they should positively update their belief about the pandemic performance of their government (and vice versa). Their belief will then inform their vote, conditioned by other factors such as the menu of alternative parties (Anderson 2000), institutions concentrating or dispersing decision making power (Powell and Whitten 1993), and political polarization based on partisanship or other salient policy issues (Kayser and Wlezien 2011). Consistent with the theory, several recent survey experimental studies

[1] Reviews on the state of the literature differ in their conclusions. A first view is that retrospective voting works pretty well at least with regard to the economy, with predictable variation across institutions (Lewis-Beck and Stegmaier 2019). A second, revisionist view is that misinformation, randomness and voter irrationality by and large limit accountability based on retrospective voting (Achen and Bartels 2016). The third view takes the middle ground that "voters sometimes, but not always, make mistakes" and argues for designing experiments to help identify behavioral biases and the conditions under which they limit the scope for accountability (Healy and Malhotra 2013: 286).

have shown that, on average, random variation in benchmarked information on the economy substantively shifts individuals' support for the government (Dassonneville and Hooghe 2016; Hansen, Olsen and Bech 2015; Olsen 2017; Tilley and Hobolt 2011).

However, in the real world individuals are exposed, at least some of the time, to different benchmarks depending on their political beliefs. With the digital revolution and growth of social media, individual choice of information is as important as ever. Thus, we extend the existing benchmarking perspective on accountability by adding the possibility of *endogenous* benchmarking. Drawing on a largely separate literature in political psychology and communication on motivated reasoning and selective news exposure (Bakshy, Messing and Adamic 2015; Kunda 1990; Lodge and Taber 2000; Taber and Lodge 2006), we argue that paying more attention to endogenous benchmarking improves our understanding of democratic accountability. The key idea is that when voters have a choice between different cross-national benchmarks, they are likely to select benchmarks that are more in line with their political orientation. Endogenous benchmarking offers a theoretical lens to further examine the conditional nature of electoral accountability depending on the supply and demand of cross-national benchmarks.

We test implications of endogenous benchmarking using pre-registered survey experiments conducted in three major European countries – France, Germany and the United Kingdom – during the COVID-19 pandemic. The pandemic constitutes an instructive test case. It threatened lives and economic wellbeing on a scale not experienced in Europe and North America since the end of World War II. In response, different governments took different policy measures and there was large variation in outcomes across countries (Engler et al. 2021). The extensive media coverage and ubiquity of cross-national

benchmarks enhance the experiments' external validity.

Building on experiments with choice protocols (e.g., Arceneaux and Johnson 2013; Gaines and Kuklinski 2011), our design combines random assignment to information treatments with a non-random assignment condition where individuals choose their preferred benchmark based on competing headlines. Importantly, assignment to the random versus non-random assignment condition is itself randomized. The design enables us to assess several empirical questions touching on key informational mechanisms enhancing or restricting accountability. First, is there evidence for endogenous benchmarking? Specifically, when given the opportunity do individuals self-select into benchmark treatments based on their prior view of the government? Second, how responsive are individuals to exogenous benchmarking information when evaluating government performance?

Our first experiment, which was conducted in the early stage of the pandemic (N=3,765), reveals clear evidence of self-selection into cross-national benchmarks consistent with motivated reasoning. In all three countries, individuals starting with a positive view of the government are much more likely to select conforming positive (for their country) rather than negative information based on the benchmarked headline. The pooled estimate suggests that a two-standard deviation increase in pre-treatment satisfaction with the government is associated with a 27 percentage point increase in the probability of choosing a positive benchmark. In a second experiment conducted during a later phase of the pandemic in one country (N=2,035), we conceptually replicate the self-selection finding for the important health policy issue of vaccinations.

We find mixed evidence for the hypothesis that individuals' evaluations of government performance during the crisis responds to additional information. While on average

participants receiving a positive benchmark become more likely to agree that their government has handled the crisis well relative to most other countries, the effect is statistically significant at the 5 percent level only in the pooled sample in the first experiment. Taken together, our results highlight the importance of political self-selection into benchmarks as a limiting factor for political accountability.

The importance of information choice for accountability goes beyond cross-national benchmarking. While self-selection into political information is not a new idea, its relevance has been hard to assess with observational data, resulting in considerable controversy (Stroud 2008). While much of the experimental work on motivated reasoning in politics focuses on biased processing of given information (Cotter, Lodge and Vidigal 2020), recent experimental studies of selective exposure in political science have found, for instance, that partisans prefer news stories that appear congenial based on the label of the news source (Iyengar and Hahn 2009; Taber and Lodge 2006). Other experiments have studied how the option to tune out of news altogether shapes opinion formation (Arceneaux and Johnson 2013). Adding to this body of research, our experiments show that individuals' political orientation predicts their choice of information even in the absence of partisan source labels, and that self-selection is evident in all countries studied and using two different designs. One implication of our findings is that individual choice of information likely matters *across* and *within* news sources and social media feeds.

This paper also speaks to the literature on differential processing of the same political information. Endogenous benchmarking is distinct from, and complementary to, accounts emphasizing that individuals exposed to the same factual information differentially attribute blame based on prior political dispositions such as partisanship (Bisgaard 2019;

Malhotra and Kuo 2008; Tilley and Hobolt 2011). In line with arguments about parallel persuasion (Coppock 2022; Wood and Porter 2019), estimates from the forced exposure conditions in our experiments suggest that, on average, individuals change their evaluations of government performance in the direction of exogenous information treatments, with no statistically significant differences in the effects across groups defined by political views or media consumption. But our main finding is that when individuals have a menu of choice, they tend to sort into different information sets based on their political orientation. This results not in "alternative facts" (e.g., about a country's vaccination rate), but rather in different benchmarks used to make sense of performance information when attributing political blame.

Endogenous benchmarking across borders

From the beginning of the COVID-19 pandemic, the World Health Organization (WHO) emphasized the importance of rapid testing of symptomatic cases for containing the spread of the virus. However, implementation of these guidelines was often lacking. For example, the British media reported that the UK struggled to implement this recommendation on a large scale. On its own, this does not necessarily imply that citizens will conclude that their government is doing a bad job. Benchmarking theories of accountability argue that evaluations depend on the yardstick used. If all similarly advanced countries countries face a test shortage, the UK's shortage is less of an indicator of bad performance than when at least some countries do better. In the former case, one may conclude that the government is not unusually incompetent or that external constraints are binding. In line with the latter case, the British media frequently contrasted testing in the UK with

Germany. For example, the UK chief medical officer was quoted stating that the UK should learn from the German example. Clearly, this benchmarked information lends itself to a less favorable evaluation of the British government.[2]

The use of benchmarking as a tool for accountability is well grounded in the political science literature on economic voting. In the clear-cut theoretical formulation of Kayser and Peress (2012), benchmarking across borders helps voters to form a judgement about how well the government has managed the macroeconomy. The media provides benchmarked information that can serve as a heuristic not just for sophisticated voters, but for a broad segment of the electorate. Recent work has formally developed a theory of reference-dependent belief formation (Aytaç 2018) and identified cross-national reference points commonly used in the media (Park 2019).[3] While there are competing interpretations of whether the available cross-national evidence supports benchmarking theories of accountability (Arel-Bundock, Blais and Dassonneville 2019; Kayser and Peress 2019; Park 2019), several experimental studies provide evidence that random variation in benchmarked information on the economy meaningfully shifts respondents' attribution of political blame (Dassonneville and Hooghe 2016; Hansen, Olsen and Bech 2015; James and Moseley 2014; Olsen 2017). Of course, benchmarks need not be cross-national, as historical or within-country comparisons are also informative (Aytaç 2018; Besley and Case 1995). However, in the pandemic studied here, contemporary cross-national comparisons were salient in the media (Krastev 2020).

In existing theoretical accounts of benchmarking and electoral accountability, as well as in related experiments, individuals are exogenously exposed to information. Studies

[2] *The Guardian*, "UK must learn from German response to Covid-19, says Whitty", 7 April, 2020.
[3] Economics has long studied yardstick competition between jurisdictions as a means to control agency problems (e.g., Besley and Case 1995).

in the literature assume (implicitly or explicitly) a relatively homogenous information environment, where individuals are exogenously exposed to benchmarks that do not vary systematically with voters' political orientation. Closely related, standard formal models of accountability—both of the selection and moral hazard variety—assume that individuals receive an exogenous performance signal (Achen and Bartels 2016).

Conceptually, we integrate the possibility of politically selective exposure into benchmarking theories of accountability. The selection mechanism may blunt the informational benefits of benchmarking. In a large literature in political psychology and behavior, theories of motivated reasoning suggest that individuals may selectively use heuristics or seek out information to justify an already held (or desired) conclusion (Kunda 1990; Taber and Lodge 2006). The result is a directional bias in information processing. While research on self-serving biases in information processing usually focuses on what information people retrieve from memory or how they process the same information (cf. Cotter, Lodge and Vidigal 2020), the logic of motivated reasoning extends to the *choice* of benchmarked information from a menu of news. The most closely related experiments look at the choice of news based on source cues in the US (Iyengar and Hahn 2009; Taber and Lodge 2006).

Endogenous benchmarking applies to individuals selectively accessing information across media sources as well as within the same source. It can take place in mainstream news sources, online or offline, or in social media news feeds. It neither requires nor implies perfect sorting into partisan echo chambers (Bakshy, Messing and Adamic 2015; Gentzkow and Shapiro 2011; Peterson, Goel and Iyengar 2021). Theory and evidence suggest that motivated reasoning may be eliminated when people have sufficient incentives to arrive at the factually correct conclusion regardless of their prior views. However, in the

context of forming political judgments in a large electorate (as well as in our experiments), these incentives are small for most ordinary people. A key observable implication of political self-selection into benchmarks is that government supporters should be more likely than opposition supporters to choose information where their country is compared favorably to a reference country.

Integrating different strands of scholarship provides a strong impetus to study the interplay between endogenous information exposure and benchmarking across borders as a tool for electoral accountability. On the one hand, benchmarked information can provide a needed input for citizens assessing their government's management of a crisis. On the other hand, self-selection shapes the benchmarks available for evaluating government performance. The extended theory suggests a conditional account of accountability. When the news media and social media provide relatively homogenous benchmarks, cross-national benchmarking enables voters to hold governments accountable. When the heterogenous supply of plausible benchmarks increases (possibly driven by individual demand in polarized times), the informational mechanism is weakened by sorting.

Endogenous benchmarking is related to, but distinct from, accounts of selective information emphasizing partisan differences in factual statements about the world (e.g., Bartels 2002). These accounts typically do not distinguish whether divergent perceptions are the result of selective processing of the same information or self-selection into different information. Our framework does not require that individuals with different political views disagree about basic facts (e.g., whether coronavirus tests are in short supply). Again, it highlights that self-selection shapes the yardstick against which governments are compared.

Experiment 1

The pandemic provides a relevant real-world setting for testing whether exogenous cross-national benchmarks affect individuals' evaluation of their government's crisis management, and, crucially, whether and how much political views shape benchmark choice.

Experimental Design

We embedded a pre-registered survey experiment in a comparative survey fielded in France, Germany, and the United Kingdom (UK) during the first wave of the COVID-19 pandemic in the spring of 2020 (see Online Appendix A.2. for the pre-registration). The pandemic is of course substantively important, but it also provides an instructive test case. While governments are not to blame for the underlying disease, different governments took different measures and outcomes varied across countries (Engler et al. 2021). The large and deadly scale of the crisis meant that individuals directly experienced its repercussions making pandemic policy highly salient.

In Europe and North America the pandemic dominated media coverage like no event since World War II. For instance, nearly one half of all stories published in the *New York Times* and *The Economist* in 2020 made reference to "covid-19" or "coronavirus" (The Economist 2020). In the month before the experiment was fielded, the pandemic was on the front-page in each issue of *The Economist*, and more than 60% of the articles mentioned the topic. The pandemic appeared no less salient in France and Germany. Political scientists also quickly noted the ubiquity of cross-national comparisons in the crisis, which meant that people were able to compare "their government's performance

with those in other countries in real time" (Krastev 2020: 54). Estimates suggest that the tone of news coverage in mainstream media was mixed rather than exclusively negative (Sacerdote, Sehgal and Cook 2020). When discussing our experimental treatments below, we provide additional examples of cross-national benchmarking by the media, some indicating that the country is doing better and others that the country is doing worse than a reference country.

In this saturated information environment, it is natural to test how individuals choose information. This is the novel part of the experiment. When assessing the impact of exogenously provided information on evaluations of how well the government is handling the crisis, we will be estimating the effect of providing *additional* information about government performance. That is, we are not examining how individuals change their views when all information is of a certain type.

Survey The survey was conducted by *Ipsos* as part of existing internet panels and was online 15-17 April 2020. The panel uses quota sampling to match the adult population in each country in terms of gender, age, occupation, region, and degree of urbanization. All estimates presented in the remainder of this paper are adjusted for sample inclusion probabilities. The dropout rate for the survey was quite low, and, more importantly, there is no evidence of item non-response related to the experiment. Table I shows sample sizes for the experiment in each country (for more survey details, see Appendix A.1.).

Experimental conditions We use a hybrid experimental design that combines exogenous treatments with self-selection to answer research questions that cannot be answered from completely randomized studies (Arceneaux and Johnson 2013; De Benedictis-Kessner

et al. 2019; Gaines and Kuklinski 2011). The experiment consists of two parts. Part *I.* provides participants with exogenously allocated positive (*a.*) or negative (*b.*) information about the pandemic in their country relative to a reference country. Part *II*. allows respondents to self-select into which information treatment they receive. Thus our design consists of three experimental conditions, in which we place respondents in each country-survey using simple random assignment. As Table I shows, we place about 25% of respondents in condition *Ia.*, 25% in condition *Ib.*, and 50% in condition *II*.[4]

Table I
Experimental groups, treatment headlines. Sample sizes in parentheses

	I. Exogenous		II. Choice
	a. positive	b. negative	
France	France takes stronger action than Great Britain	France lags behind Germany in testing	a. or .b
(1515)	*(403)*	*(404)*	*(708)*
Germany	Germany is European testing champion	Germany is laggard in acquiring masks	a. or .b
(1500)	*(399)*	*(401)*	*(700)*
United Kingdom	UK takes more forceful action than the Dutch	UK testing lags behind Germany	a. or .b
(750)	*(200)*	*(200)*	*(350)*

Note: Reference countries for Germany in vignette text are South Korea (negative) and France (positive). The complete vignette text is available in Online Appendix A.3.1.

In the exogenous benchmarking conditions, respondents are presented with vignettes in the style of a short news article. It consists of a headline, as displayed in Table I, and body text of about 70 to 80 words providing benchmarked information. Respondents are instructed to read the short text and answer the subsequent questions. For example, in the

[4]The experimental sample consists of 75% of the survey sample, as one group of respondents was allocated to not participate in the experiment, in order to have a respondent subset not exposed for the purpose of analyzing survey items not part of this experiment.

UK, respondents in group *Ia.* are presented with a headline stating that the UK took more forceful actions than the Dutch. The body text of the vignette discusses the measures taken by UK and Dutch governments. It emphasizes that "the UK has enacted a stricter lockdown" and points out that "[w]hile both countries have seen an increase in deaths from Covid-19, the Netherlands have experienced about 20 percent more deaths per 100,000 inhabitants." Respondents in group *Ib.* instead are confronted with a headline stating that the UK lags behind Germany in testing for the coronavirus. The vignette body states the WHO's recommendation for wide-spread testing in order to better control the virus and protect a country's populations. The text then quotes the government's chief medical officers admitting that the UK government has fallen behind Germany in testing.[5]

All vignettes compare a respondent's country to a reference country. This captures the fact that during the pandemic news articles often made international comparisons to one or a few comparison countries. The choice of reference countries is in line with prior research that identifies reference points based on an analysis of media coverage of economic news. Specifically, our vignettes include common reference countries identified by Park (2019) for the closest available year. For example, one headline in the *The Guardian* is that "UK must learn from German response to Covid-19, says Whitty".[6] The experiment does not employ deception. The information provided is based on facts that are credibly publicly available; quoted statements from government officials are taken

[5]Agency models with asymmetric information illustrate that more voter information does not always improve voter welfare (Ashworth and Bueno de Mesquita 2014). For instance, voters learning that a politician is a bad type can undermine the politician's incentives to work hard as there will be no re-election in equilibrium. Our focus is on the type of information, related to comparative policy responses, rather than politicians' type, that is theoretically linked to better accountability.

[6]A partial exception is Germany, where we use South Korea as a reference point in the negative vignette. This reflects the media attention given to South Korea, which was hit earlier by the crisis and took aggressive measures to flatten the curve. For example: *Tagesschau*, "South Korea as Role Model?" (our translation), 31 March, 2020.

from official news sources. The average difference in word length between positive and negative conditions is 3 words. The full text for all vignettes is available in Online Appendix A.3.1. There we also show that respondents' positively rated the quality of the vignettes across countries (see Figure A.2).

Respondents randomized into condition *II.* are able to self-select their treatment. They are presented with positive and negative benchmark headlines *a.* and *b.* and are asked to choose one of them to read the story. After choosing a headline, respondents are presented with the corresponding full vignette. Both headlines and vignette text are identical to those received by respondents in the exogenous information condition. In the second experiment, we consider a different choice setting where people are also offered a neutral headline.

The choice condition captures the fact that for salient topics like the COVID-19 pandemic, individuals often have a choice between different news reports of the same issue, both within and across media outlets, as well as on social media. For example, the British media reported that the UK was doing worse on Coronavirus testing compared to Germany, while it also reported the positive news of declining infections rates in the UK[7] and pointed out lack of large-scale testing in Germany.[8] Similarly, on the same day a leading French newspaper published two divergent articles about the progress of vaccination.[9] More broadly, a study of news coverage during the pandemic estimates that the tone of news coverage in major non-US media outlets was negative in 54 per cent of the stories and positive in 46 per cent (Sacerdote, Sehgal and Cook 2020). Relatedly, the largest

[7] *BBC*, "Coronavirus: UK cases 'could be moving in the right direction'", 7 April, 2020.
[8] *The Guardian*, "Germany told it needs to massively increase coronavirus testing'", 2 April, 2020.
[9] *Le Figaro*, "Vaccination Covid19: What is the position of France"; " 'The Slowness' of Kundera and the incredible delay of vaccination in France" (our translation). Both 5 January , 2021.

online news sites tend to be neutral in terms of partisanship (Gentzkow and Shapiro 2011). On social media most individuals are exposed to news feeds that entail a choice of information (Bakshy, Messing and Adamic 2015).[10] Thus, all vignette headlines are designed to provide no partisan cues in order to provide a stricter test of self-selection (and because such cues are not generally present in mainstream media).

Outcome variables and hypotheses Our first outcome variable is an individual's overall assessment of how well the government has responded to the pandemic. Respondents were prompted to indicate how much they agree or disagree with the statement "all in all, the government has handled Coronavirus better than most other countries?" using an 11-point scale with labelled endpoints ranging from 0 ("strongly disagree") to 10 ("strongly agree"). In line with benchmarking theory, this captures respondents' global assessment of how well the government in their country has managed the crisis. Note that this item does not immediately follow the treatment, but is placed after a battery of items asking respondents to evaluate the quality of the text, in order to reduce experimenter demand effects. Based on the discussion in the previous section, our first pre-registered hypothesis concerns the effect of exogenous information on individuals' evaluation of government performance:

Hypothesis 1 *Exposure to positive benchmarking information leads to more favorable evaluation of government performance compared to exposure to negative benchmarks, all else equal.*

[10] In Austria we fielded a different experiment: All respondents choose between competing headlines; conditional on the headline choice, there also is a light information treatment. Again, we find political sorting based on pre-treatment satisfaction with the government. Due to space constraints, results are reported in Online Appendix A.3.10.

This *exogenous benchmarking* hypothesis is based on standard benchmarking theory (Aytaç 2018; Kayser and Peress 2012; Powell and Whitten 1993), in which benchmarking across borders works as a heuristic. But it is by no means a foregone conclusion that the data reject the null hypothesis of no treatment effect. We are conducting a demanding test of the benchmarking mechanism, in the sense that the treatment concerns a comparison of a respondent's home country with another reference country, whereas the outcome variable is an assessment of the government's crisis management *in toto*. Our outcome variable is not a restatement of the fact (e.g., whether the UK tested less than Germany) but a summary political evaluation. Furthermore, the literature suggests that selective perception or interpretation can limit treatment effects. For example, heterogeneity in political predispositions may lead to divergent inferences about how well the government has dealt with an issue even when individuals agree on the facts (Bisgaard 2019; Tilley and Hobolt 2011), resulting in a null effect on average.

Our second outcome variable concerns the choice of benchmarking headline in the experimental selection condition (*II*). It enables us to test our second hypothesis, which is derived from the extended endogenous benchmarking framework. The logic of self-selection implies that individuals in the choice condition do not randomly select one of the headlines and, more specifically, that there is sorting based on pre-treatment political attitudes. We registered the use of a pre-treatment measure of *satisfaction with the government* (more precisely, the current head of the executive, referring to President Macron in France, Chancellor Merkel in Germany, and Prime Minister Jonson in the UK) on an 11-point scale ranging from "completely dissatisfied" to "completely satisfied".[11]

[11]The exact question wording is: "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")

This is an omnibus measure of political dispositions tapping into partisanship, valence and other prior evaluations of the government. Thus, the *endogenous benchmarking* hypothesis can be stated as follows:

Hypothesis 2 *Existing satisfaction with the government increases the probability of self-selecting into positive benchmarking information, all else equal.*

The design of this experiment is not meant to examine whether information using a reference country works differently than information using history or no reference point at all. Prior experimental studies (focused on the economy) have shown the effectiveness of exogenous benchmarking in this regard (Dassonneville and Hooghe 2016; Hansen, Olsen and Bech 2015; Olsen 2017; Tilley and Hobolt 2011). Rather, it is designed to analyze whether individuals are responsive to exogenous information during the pandemic and, going beyond previous work, estimate the relevance of self-selection into alternative benchmarks.

Background variables To analyze effect heterogeneity when examining the exogenous benchmarking hypothesis, we use pre-treatment measures of media usage, trust in the media, satisfaction with democracy, as well as the satisfaction with the chief executive discussed above.[12] *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. We capture *trust in the media* by inviting respondents to indicate how much they trust journalists on a labelled 4-point scale ranging from "trust completely" to "don't trust at all". We measure *satisfaction with democracy* using a standard item on an 11-point

to 5 ("neither nor") to 10 ("completely satisfied").

[12] See Online Appendix A.3.2. for details.

rating scale ranging from “not satisfied at all” to “completely satisfied”.

Main Results

Endogenous benchmarking

In a diverse media environment and even within the same media outlet during a multi-dimensional crisis, individuals often have the choice of what cross-national benchmark they consider when evaluating their country’s performance on a salient issue. The endogenous benchmarking hypothesis (H2) concerns the choice of benchmarks based on prior political dispositions. Analyzing choice condition *II* in the experiment, we are able to assess the empirical relevance of self-selection. We find clear evidence that individuals purposefully select into receiving specific benchmarking headlines.

Descriptively, the overall pattern of survey participants’ choices deviates significantly from what one would expect to observe if they simply chose a headline at random. The final column of Table II shows p-values from an exact test comparing observed proportions to the null hypothesis of a binomial distribution with probability parameter 0.5. In all countries, the null hypothesis of a 0.5 proportion is rejected. This pattern is also evident by simply looking at the observed proportion of respondents who selected positive benchmark headlines. Roughly two thirds of respondents chose a *negative* headline, while about one third chose to receive a positive benchmark (there is no item nonresponse at this stage). This indicates a tendency of respondents to seek out critical information during the COVID-19 pandemic. This is in line with results from social psychological experiments indicating that negative stimuli attract more attention and are more likely to be selected (Fiske 1980), which may be seen as more informative and diagnostic or due a general

tendency towards negativity in the political arena.

Table II
Exact Binomial test of non-random benchmark selection.

	Proportion positive	$H_0 : Pr = 0.5$ p-value
Pooled sample	0.310	0.000
France	0.310	0.000
Germany	0.330	0.000
United Kingdom	0.290	0.000

Note: Exact two-sided test of proportion using as null distribution the Binomial distribution with parameter 0.5.

Does a pro-government predisposition determine the choice between two competing headlines? Our specific hypothesis is that self-selection is related a respondent's pre-treatment satisfaction with the government in general. Figure I plots the estimated association between respondents' pre-treatment political orientation and their propensity to choose the positive (for their country) benchmark headline. The left panel uses satisfaction with actions of the chief executive (as specified in the pre-analysis plan), while the right panel uses party identification to capture individuals' prior political orientations.[13] Partisanship is an indicator variable equal to one if a respondent identifies with the governing party (i.e., the party of the chief executive). Based on both measures we find clear evidence of a systematic relationship between respondents' prior views and their information choice in all three countries. Adjusting for pre-treatment covariates barely changes the estimates.[14]

Respondents who were more satisfied with their government leader prior to the experi-

[13] We thank an anonymous reviewer for pointing us towards this additional analysis.
[14] Pre-treatment covariates are age in years, indicators for female, college education, and employment status.

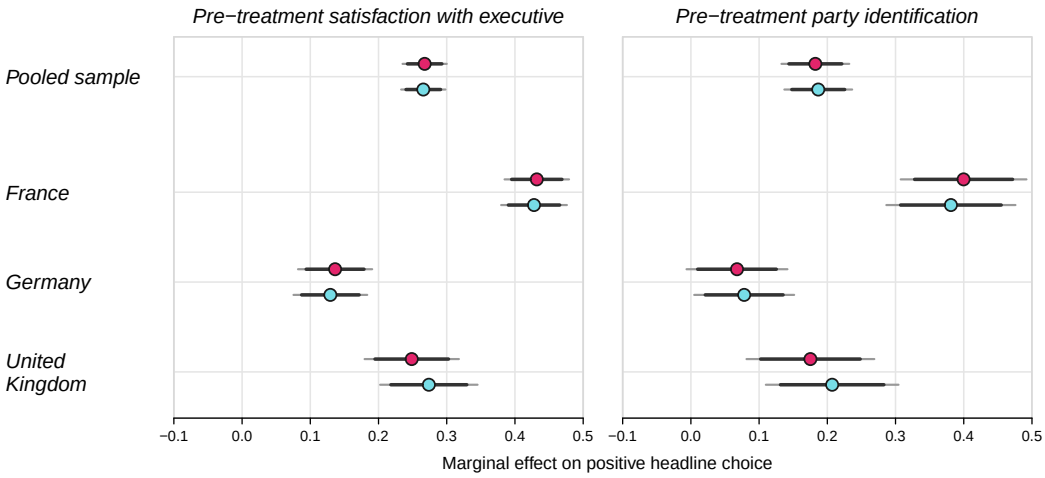

Figure I
Pre-treatment political orientation and positive benchmark selection

Note: Marginal effects of pre-treatment satisfaction with head of executive and pre-treatment party identification (indicator variable for identifying with the governing party) on the probability of a respondent choosing a positive cross-national benchmark (for the country). Shown are marginal effects calculated from linear probability models without covariates (●) and adjusted (●) for survey-design (pre-treatment) covariates. Satisfaction is scaled by two standard deviations (Gelman 2008). Confidence intervals (with 90% and 95% coverage) are based on heteroscedasticity-consistent standard errors.

ment were more likely to choose the headline that makes their country's performance look good compared to a reference country on some dimension of the pandemic. On average in the pooled model, a two-standard deviation (SD) increase in prior satisfaction is associated with a 27 percentage point increase in the probability of choosing a positive benchmark. This relationship most pronounced in France and least pronounced in Germany (where the marginal effect is about 14 points). The relationship in the UK resembles the pooled sample estimate. However, even in Germany the association is statistically significant and substantively meaningful.[15] To provide another view on the substantive magnitude of this effect, we calculate first differences in choice probabilities

[15]The mean of pre-treatment satisfaction is 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) though it is lower in France (4.2). In France there are more people who are completely dissatisfied with their government (see Figure A.1). The difference might explain why the marginal effect is largest in France but not why it is larger in the UK than in Germany.

when shifting a respondent with a median level of satisfaction to the 90th percentile. The probability of choosing a positive headline increases by 17.9 percentage points in the pooled sample (*s.e.*=1.4), by 12.2 (*s.e.*=2.3) and 23.2 (*s.e.*=1.6) percentage points in the UK and France, respectively, and by 6.9 (*s.e.*=1.8) points in Germany. Still, self-selection is not complete. Even among government supporters, a significant number of individuals preferred negative news. Among opponents of the government, a smaller but non-trivial number of individuals searched out positive news (see Online Appendix Figure A.3).

We find a similarly clear relationship when using party identification to measure political orientation. As shown in the right panel of Figure I, in a pooled analysis, individuals who identify with the governing party are 19 percentage points more likely to choose the positive benchmark compared to those who do not identify with the governing party. In single country analyses, the largest effect appears in France (38 percentage points), while the UK estimate is close to the pooled one. The estimate in Germany is again the smallest (about 7.8 percentage points).[16]

The estimates show that individuals' overall political orientation is strongly associated with their choice of information in the experiment. These results are consistent with motivated reasoning (e.g., Lodge and Taber 2000; Taber and Lodge 2006). An alternative interpretation might be that individuals are accuracy-seeking and use headlines as cues about which source might be more credible given their prior disposition (Druckman and McGrath 2019). While more nuanced, this argument implies the same result for accountability: individuals choose benchmarks aligned with their political predispositions.

[16]Estimates for Germany, where the coalition government includes the two largest parties, are the same when measuring partisanship as alignment with either of the two parties in the coalition government. Relatedly, one intriguing possibility is that in Germany joint decision-making between the federal government and state governments blurs political responsibility and thereby dampens the motivation for directional information choice. This is, however, beyond the scope (and capability) of this paper.

While it is not easy to distinguish the mechanisms empirically, we find the latter possibility less plausible. In the experiment, self-selection emerges despite the absence of explicit source cues in the competing headlines. The design constitutes a harder test for political sorting. It is also worth noting that differences in the perceived credibility of the vignette across exogenous and endogenous benchmarks (see Online Appendix Figure A.2) are minute compared to the magnitude of the political self-selection effect in headline choices shown in Figure I.

The political bias in the benchmark selection uncovered here is also not easily accounted for by Bayesian learning. In the foundational Bayesian learning model, the signal is exogenous (Bullock 2009). Bayesian models with information choice often focus on attention as a scarce resource (Matějka and Tabellini 2020). These models do not predict that individuals should choose information in line with their political leanings. To be clear, the experiment does not aim to test a Bayesian model with information choice. This would require a different design. Rather, the findings highlight a neglected aspect of partisan information processing that has implications for the demand side of information bearing on accountability. By screening out countervailing information, self-selection weakens the informational chain of accountability.

Exogenous provision of benchmarking information

What if individuals are exogenously exposed to benchmarking information, as they are in prior studies? Based on the forced exposure part of the experiment, Figure II summarizes the main results concerning the effect of exogenously provided information on public evaluations of the government's response to the pandemic based on experimental conditions *Ia* and *Ib*. For each country as well as the pooled sample, it plots the average

treatment effect of providing the positive cross-national comparison versus the negative cross-national one based on difference-in-means and covariate-adjusted estimates.[17]

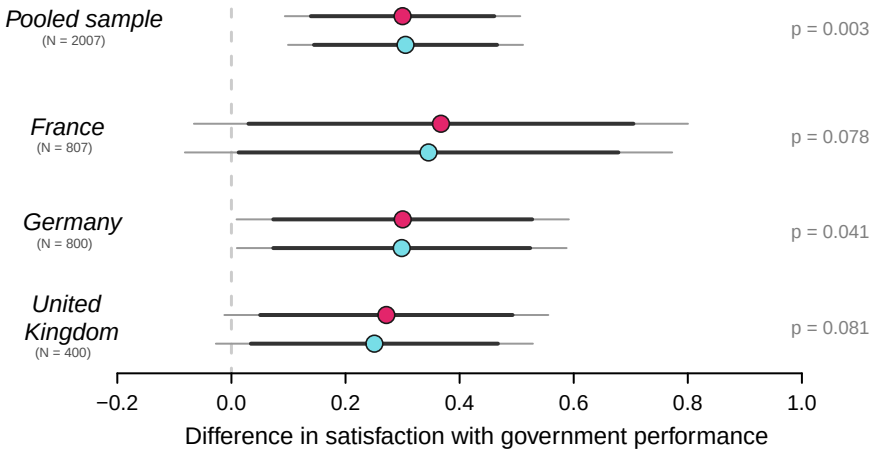

Figure II
Exogenous information and evaluation of government performance

Note: Average treatment effects of exogenous provision of positive versus negative benchmarking information. Difference-in-means (●) and covariate-adjusted (○) estimates. Confidence intervals (with 90% and 95% coverage) are based on heteroscedasticity-consistent standard errors. Randomization *p*-values testing sharp directional null hypothesis shown on the far right.

The estimates show that the exogenous information treatments tend, on average, to move respondents' views on how well the government has handled the pandemic. In the pooled sample, the average treatment effect is 0.30 units on the 11-point scale (*s.e.*=0.13). The direction of the effect of exogenous benchmarks on individuals' overall evaluation of the government is in line with the standard benchmarking theory assuming exogenous information provision (Aytaç 2018; Kayser and Peress 2012). Respondents who receive information that makes their own country look good compared to a comparison country have more positive evaluations of their government's management of the crisis compared to most other countries. Statistically, in the pooled model we can reject the null hypothesis

[17]When adjusting for pre-treatment covariates (age in years, indicators for female, college education, and being employed), we follow the setup of Lin (2013).

of no effect at the five percent level (whether one uses asymptotic or randomization p-values). The estimates are practically identical across estimation methods (adjusted or unadjusted for covariates). While estimates in the country samples are more uncertain, they all have the same sign and are rather similar (and "statistically significant" if one is prepared to employ a more generous $p < 0.1$ threshold).[18]

Assessing the substantive magnitude of the effect is somewhat more subjective. The average effect of the positive cross-national benchmark of 0.3 points (in the pooled model) represents a 1/10th standard deviation shift of the dependent variable. When compared to average evaluations in the experimental group receiving the negative benchmark (4.96), this effect amounts to a 6 per cent increase (see Online Appendix Table A.3 for effect sizes expressed in terms of standard deviations and percentages in individual countries and with covariate adjustment; Table A.2 provides detailed descriptive statistics). The effect is roughly similar to the effect of cross-national benchmarking on the economy in a related choice experiment conducted in Denmark (Hansen, Olsen and Bech 2015: 783). Given that information on government performance in the pandemic was plentiful, one would not necessarily expect that a *single* benchmark completely changes an individual's global view of the government. Bayesian and sampling models of information processing imply a positive but declining marginal effect of additional signals in such an environment. Altogether, it is fair to say that the effect of exogenous information seems modest.[19]

In additional analyses reported in the Online Appendix, we explore the heterogeneity of the information effect from the forced exposure. Average effects can hide differential

[18] Unlike France and the UK, the German headlines do not mention the reference country. This does not affect the estimated treatment effect (Online Appendix Table A.6).

[19] In Online Appendix A.3.9 we study the impact of benchmarking information and performance evaluations on vote choice as a more distal outcome. We find that the exogenous benchmarking treatments affect vote intention through comparative evaluations.

responses according to characteristics like prior satisfaction with the government, satisfaction with democracy, media usage, or trust in the media. However, we fail to reject the null hypothesis of no heterogeneity across the pre-specified variables (Online Appendix A.3.7). This also implies no evidence of backlash against non-congruent information (Coppock 2022; Wood and Porter 2019).

Experiment 2

The second experiment serves two purposes. First, we test if self-selection occurs in a later stage of the pandemic, in which a different policy—vaccinations—has become the central issue. We also offer individuals a neutral headline and present benchmarking information in a more quantitative fashion (via a tabular comparison). Second, we employ a different design in order to analyze the impact of new benchmarking information *after* self-selection. This follow-up experiment was conducted in France as part of the same *Ipsos* internet panel used for the first experiment. It was in the field during the third wave of the pandemic, on 11-13 March, 2021, with a sample size of 2,035.

As is illustrated in Figure III, experiment 2 uses a three-stage design. All respondents face an information choice at the second stage (II). The first stage (I) randomizes the choice set. Based on the initial random assignment, half of the sample is asked to choose between a story on vaccinations with a neutral headline ("Is France doing better or worse?") and a headline indicating positive content ("France far from being at the back of the pack"). The other half of the sample is asked to choose between a story based on the same neutral headline and a headline indicating negative content ("France far from the best"). The choice part of the experiment enables us to test for the relevance

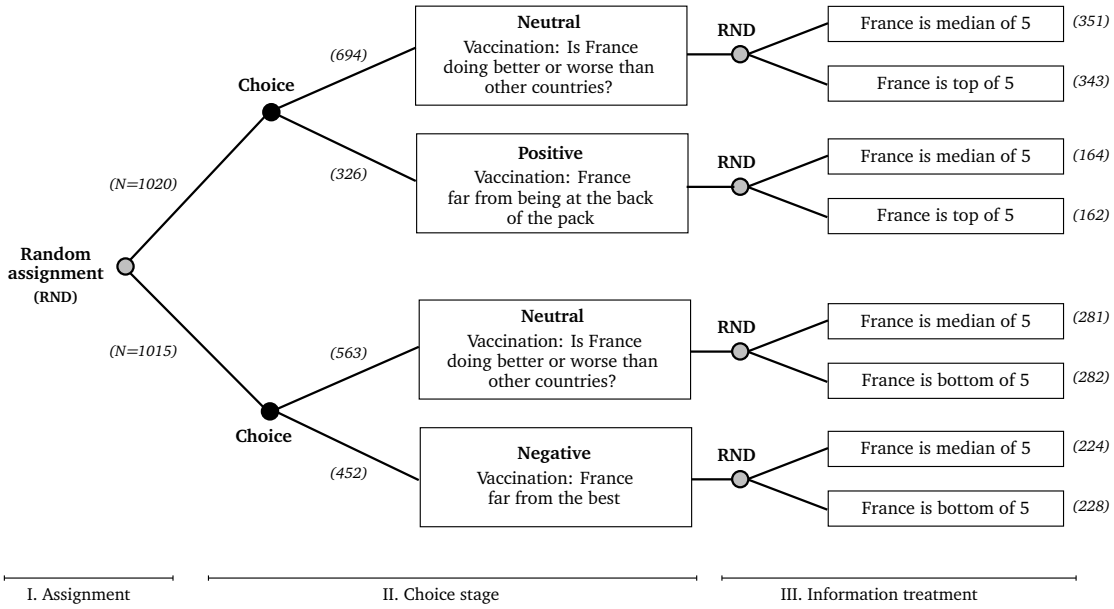

Figure III

Experiment 2: Three-stage design. Respondent choices and randomized benchmarks.

Note: Number of observations in parentheses. The complete vignette text and the list of five comparison countries is available in Online Appendix A.4.1.

of endogenous benchmarking in a different environment. In contrast to experiment 1, the choice is less sharp. The comparison is no longer between a positive and a negative headline. Rather, it concerns the choice between a neutral and a positive or between a neutral and a negative one. Moreover, the information choice focuses on a different aspect of the pandemic: vaccinations. We assess whether political motivations still drive self-selection. Given the experimental design, the self-selection hypothesis implies that pre-treatment satisfaction with the government increases the probability of choosing a positive versus a neutral and a neutral rather versus a negative headline.

The final information stage (III) provides respondents with detailed benchmarking information based on a ranking of five countries. We use simple random assignment to either display positive or neutral information (for respondents in the first group) or

negative or neutral information (for those in the second group). Another reason for the initial randomization into two groups—one choosing between neutral and positive, the other between neutral and negative—is to allow for the randomization of benchmarking information in stage III consistent with each headline.[20]

Any given respondent sees one of three vignettes. Each vignette has the same introductory text stating that the campaign to vaccinate people against the coronavirus has begun several months ago and asks how well the respondent's country is doing compared to other countries (exact wording is available in Appendix A.4.1). This text is accompanied by a compact table that shows quantitatively how France compares to four other OECD countries in terms of the percentage of individuals vaccinated so far. The information provided is factually correct. The experimental variation in the vignette consists in the choice of benchmark countries included in the comparative table. In the neutral benchmarking treatment, France is the median country out of five countries, including a vaccination leader (UK), a vaccination laggard (Australia) and two neighboring countries with similar vaccination rates (Belgium and Germany). In the positive information treatment, France is compared favorably to four countries with lower vaccination rates (Canada, Austria, South Korea and Australia). In the negative treatment, France is compared unfavorably to four countries with higher vaccination rates (US, UK, Denmark, Spain).

What is the effect of exogenous benchmarking across borders on vaccinations conditional on a prior choice of a neutral or directional headline? With respect to government accountability, our main outcome variable is the same as in the previous experiment: respondents' overall assessment of how well the government has responded to the pandemic

[20]The setup for analyzing heterogeneity based on self-selection differs from the design by Gaines and Kuklinski (2011), which uses a principal stratification approach.

on a 11-point scale. The experiment captures that while individuals may try to select congenial information based on cues like a headline, they do not control the fuller information they receive once they read a story. For instance, a person seeking out negative news may receive information that France is in the middle of the pack rather than at the bottom in terms of vaccinations. Following standard benchmarking theory, the exogenous benchmarking hypothesis is that there should be a negative (positive) marginal effect of seeing France ranked bottom (top) rather than in the middle, regardless of whether people initially selected a neutral or directional headline. In addition, the experiment enables us to asses if information effects vary across self-selected groups. Our first experiment did not find much heterogeneity based on observable pre-treatment characteristics. Going further, this experiment enables us to directly condition on the choice of the benchmarking headline. One conjecture is that individuals who are more eager to reach a particular conclusion, as revealed by their choice of a directional headline, may be less receptive to opposing information.

Results

Experiment 2 yields clear evidence in support of endogenous benchmarking, bolstering the results from the first experiment. Figure IV shows that strong supporters of the government are significantly more likely to choose a positive over a neutral headline. A two SD increase in pre-treatment satisfaction with the government is associated with a 20 percentage point increase in the probability of positive benchmark selection.[21] Similarly, when facing the choice between a negative and a neutral headline, a two SD increase in

[21] We scale satisfaction to two SDs for consistency with Figure I. Online Appendix A.4.2 provides further details and estimates.

pre-treatment satisfaction with the government is associated with a 35 percentage point decrease in the probability of selecting a negative benchmark.

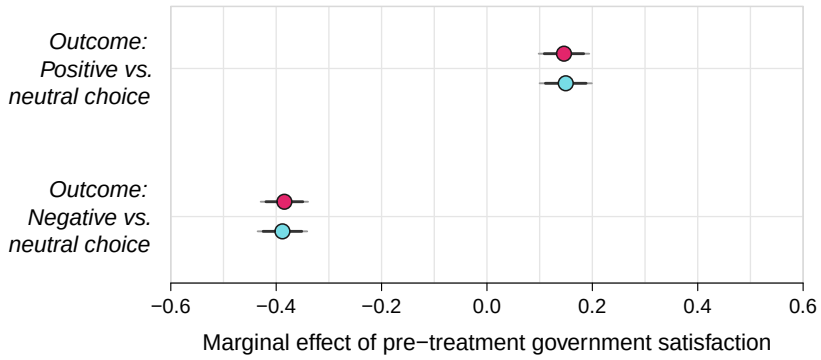

Figure IV
Pre-treatment political orientation and benchmark selection

Note: Marginal effects of pre-treatment satisfaction with head of executive on the probability of a respondent choosing a (i) positive vs. neutral or (ii) negative vs. neutral benchmark in France. Shown are marginal effects calculated from linear probability models without covariates (●) and adjusted (●) for survey-design (pre-treatment) covariates. Confidence intervals (90% and 95%) are based on robust standard errors.

To provide another perspective on the substantive impact of endogenous benchmark choice, we can calculate the change in choice probability when moving a respondent from the median levels of satisfaction to the 90th percentile of the satisfaction distribution. This shift increases the probability of choosing a positive benchmark by about 10 percentage points, while it decreases the probability of choosing a negative benchmark by 27 points.

Next, we turn to analyzing the link between exogenous benchmarking and global performance evaluations for different self-selected types of respondents. Figure V displays the resulting estimates of the average treatment effects, all weighted by sample inclusion probabilities, with confidence intervals based on robust standard errors. The two estimates at the bottom of Figure V are from the group who, at stage II, had the choice between a neutral and a positive headline. The estimates indicate that receiving the positive benchmark ('France is top of 5') rather than the neutral one ('France is median') in stage

III of the experiment has essentially no impact on performance evaluations. The difference estimate is close to zero and the confidence intervals are wide. This holds regardless of respondents' revealed type, that is, whether they have previously chosen a positive (black estimate) or neutral (light gray estimate) headline. Thus, heterogeneity of the treatment effect across self-selected groups is negligible.

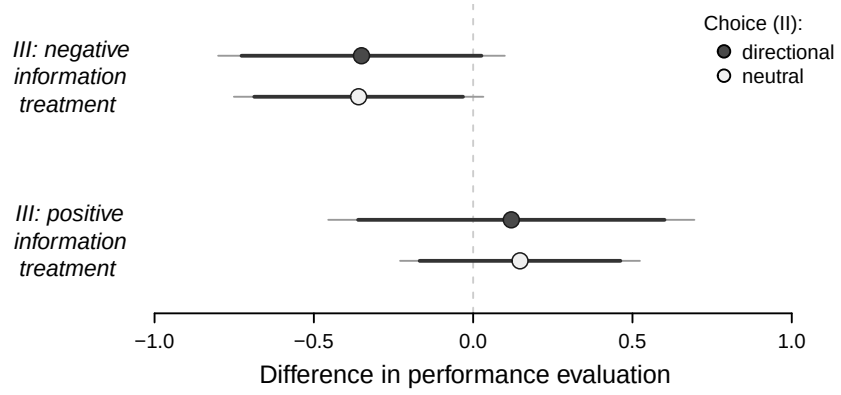

Figure V
Benchmark choice, exogenous benchmarking information, and evaluation of government performance.

Note: Shown are group differences weighted by sample inclusion probability. Confidence intervals (with 90% and 95% coverage) are based on robust standard errors.

The two estimates at the top of Figure V are based on the second experimental group, in which self-selection is based on the choice (at stage II) between a neutral and a negative headline. We find a somewhat larger difference in average evaluations between the benchmark treatments. For neutral-choosers exposed to the negative benchmark, evaluations drop by 0.36 points (compared to the neutral benchmark). The magnitude of this difference is very similar to the effect of the exogenous information treatment estimated in the first experiment. However, note that the confidence intervals are rather wide, rendering the estimate statistically insignificant at the 5% level (this also holds when adjusting for covariates; cf. Table A.9). For individuals that have chose the negative

headline in stage II, the difference in performance evaluations between the randomized benchmarks is virtually identical to the neutral types (0.35 points).[22]

The findings provide little additional support for the exogenous benchmarking hypothesis. The estimates for the exogenous benchmarking treatments conditional on prior self-selection are close to zero or, when they are larger, come with relatively wide confidence intervals. The estimates of randomized information are also rather homogenous across self-selected groups, consistent with the limited heterogeneity found in experiment 1. Taken together, our results highlight the importance of accounting for prior self-selection into information as a mechanism limiting political accountability.

Conclusion

While cross-national comparisons are a powerful source of accountability in modern democracies (Kayser and Peress 2012), endogenous benchmarking can weaken it. The survey experiments we conducted in three countries during the worst pandemic in a century demonstrate that when given the opportunity to choose, individuals systematically self-select into benchmarks in line with their prior (ideological) view of the government. While selection effects play a central role in other literatures, they received little attention in previous work on benchmarking across borders and accountability. Going beyond other recent work on motivated reasoning and information choice in political science (Iyengar and Hahn 2009; Taber and Lodge 2006), self-selection emerged in our experiments despite

[22] Why is there a larger difference between benchmark treatments in the second group than in the first? One potential explanation is some form of "last-place aversion" (Kuziemko et al. 2014; Zhou and Soman 2003): individuals are more averse to their country being at the bottom of the table than to being in the middle versus the top. An alternative conjecture is related to the information environment discussed above: respondents in the neutral-positive group might be aware that while France was quicker in vaccinating its population than some OECD countries it was not part of the vaccination vanguard.

the absence of strong source cues in all countries and using two different experimental designs.

The experiments were conducted in a global crises that received substantial media attention and where heterogenous benchmarks were common. In this setting, only looking at the impact of exogenously varied benchmarks risks substantively overstating the informational benefits of cross-national benchmarking. Endogenous benchmarking implies that not everybody will be exposed to the same information. In other situations, individuals may face a homogenous set of comparison cases. When the supply of benchmarks is more homogenous, there is less scope for political self-selection and benchmarking across borders becomes effectively exogenous for many voters. One important avenue for future work is to examine the political supply and variation in benchmarks across issues and over time (extending work by Park 2019). Relatedly, a promising extension of our experiment would be to expand the set of available options in the choice condition, for example by including pure entertainment as an option (Arceneaux and Johnson 2013).

Supplementary Material

The Online Appendix with supplementary material is made available on the Cambridge University platform alongside the article. It contains: A.1 Survey Details; A.2 Pre-registration; A.3 Experiment 1 (vignette, wording, wording of survey questions, descriptive statistics, respondent evaluations of texts, additional analyses of benchmark choice, estimates of exogenous benchmarking, treatment effect heterogeneity, vote choice, additional experiment: Austria); A.4 Experiment 2 (vignette wording, additional results).

Data Availability Statement

The data, replication code, and codebook (Becher, Brouard and Stegmueller 2023) can be found at `https://doi.org/10.7910/DVN/BY1SN7`

Acknowledgements

For comments and suggestions on an earlier version, we are especially grateful to our three anonymous reviewers, Kevin Arceneaux, Zuheir Desai, Miriam Golden, Macarthan Humphreys, Peter John, Moritz Marbach, Karine Van Der Straeten, participants in (virtual) seminars at Berlin Social Science Center (WZB), IAST, IE University, EPSA 2020, European University Institute, MPSA 2021, Sciences Po, and Texas A&M. Stefan Preuß provided excellent research assistance in the first experiment.

Financial Support

Becher acknowledges financial support from IE University and IAST funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010. Sylvain Brouard acknowledges the financial support from ANR–REPEAT grant (Special COVID-19), CNRS, Fondation de l'innovation politique, as well as regions Nouvelle-Aquitaine and Occitanie. Stegmueller acknowledges funding from Duke University and the National Research Foundation of Korea (NRF-2017S1A3A2066657).

Competing Interests

None.

Ethical Standards

The research was conducted in accordance with the protocols approved by the Review Board for Ethical Standards in Research at the Toulouse School of Economics and the Institute for Advanced Study (ref.code 2020-04-001).

References

Achen, Christopher H. and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton and Oxford: Princeton University Press.

Anderson, Christopher J. 2000. "Economic voting and political context:a comparative perspective." *Electoral Studies* 19(2-3):151–170.

Arceneaux, Kevin and Martin Johnson. 2013. *Changing Minds or Changing Channels? Partisan News in an Age of Choice*. Chicago: University of Chicago Press.

Arel-Bundock, Vincent, André Blais and Ruth Dassonneville. 2019. "Do Voters Benchmark Economic Performance?" *British Journal of Political Science* 51:437–449.

Ashworth, Scott and Ethan Bueno de Mesquita. 2014. "Is Voter Competence Good for Voters?: Information, Rationality, and Democratic Performance." *The American Political Science Review* 108(3):565–587.

Aytaç, Selim Erdem. 2018. "Relative Economic Performance and the Incumbent Vote: A Reference Point Theory." *The Journal of Politics* 80(1):16–29.

Bakshy, Eytan, Solomon Messing and Lada A. Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348(6239):1130–1132.

Bartels, Larry M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Becher, Michael, Sylvain Brouard and Daniel Stegmueller. 2023. "Replication Data for: Endogenous benchmarking and government accountability: Experimental evidence from the COVID-19 pandemic." `https://doi.org/10.7910/DVN/BY1SN7`, Harvard Dataverse.

Besley, Timothy and Anne Case. 1995. "Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition." *The American Economic Review* 85(1):25–45.

Bisgaard, Martin. 2019. "How Getting the Facts Right Can Fuel Partisan-Motivated Reasoning." *American Journal of Political Science* 63(4):824–839.

Bullock, John G. 2009. "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion." *The Journal of Politics* 71(3):1109–1124.

Coppock, Alexander. 2022. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago: University of Chicago Press.

Cotter, Ryan G., Milton Lodge and Robert Vidigal. 2020. When, How, and Why Persuasion Fails: A Motivated Reasoning Account. In *The Oxford Handbook of Electoral Persuasion*, ed. Elizabeth Suhay, Bernard Grofman and Alexander H. Trechsel. Oxford University Press pp. 51–65.

Dassonneville, Ruth and Marc Hooghe. 2016. "Are Voters Benchmarking the National Economy? An Experimental Test During the 2014 US Congressional Elections." Paper presented at the Annual Conference of the Canadian Political Science Association (CPSA), University of Calgary, May 31 - June 2, 2016.

De Benedictis-Kessner, Justin, Matthew A. Baum, Adam Berinski and Teppei Yamamoto.

2019. “Persuading the Enemy: Estimating the Persuasive Effects of Partisan Media with the Preference-Incorporating Choice and Assignment Design.” *American Political Science Review* 113(4):902–916.

Druckman, James N. and Mary C. McGrath. 2019. “The evidence for motivated reasoning in climate change preference formation.” *Nature Climate Change* 9(2):111–119.

Engler, Sarah, Palmo Brunner, Romane Loviat, Tarik Abou-Chadi, Lucas Leemann, Andreas Glaser and Daniel Kübler. 2021. “Democracy in times of the pandemic: explaining the variation of COVID-19 policies across European democracies.” *West European Politics* 44(5-6):1077–1102.

Fiske, Susan T. 1980. “Attention and weight in person perception: The impact of negative and extreme behavior.” *Journal of Personality and Social Psychology* 38(6):889–906.

Gaines, Brian J. and James H. Kuklinski. 2011. “Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection.” *American Journal of Political Science* 55(3):724–736.

Gelman, Andrew. 2008. “Scaling regression inputs by dividing by two standard deviations.” *Statistics in Medicine* (October 2007):2865–2873.

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

Hansen, Kasper M., Asmus L. Olsen and Mickael Bech. 2015. “Cross-National Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic.” *Political Behavior* 37:767–789.

Healy, Andrew and Neil Malhotra. 2013. “Retrospective Voting Reconsidered.” *Annual Review of Political Science* 16(1):285–306.

Iyengar, Shanto and Kyu S Hahn. 2009. “Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use.” *Journal of Communication* 59(1):19–39.

James, Oliver and Alice Moseley. 2014. “Does performance information about public services affect citizens’ perceptions, satisfaction and voice behavior? Field experiments with absolute and relative performance information.” *Public Administration* 92(2):493–511.

Kayser, Mark A. and Michael Peress. 2012. “Benchmarking across Borders: Electoral Accountability and the Necessity of Comparison.” *American Political Science Review* 106(3):661–684.

Kayser, Mark A. and Michael Peress. 2019. “Benchmarking Across Borders: An Update and Response.” *British Journal of Political Science* 51:450–453.

Kayser, Mark Andrea and Christopher Wlezien. 2011. “Performance pressure: Patterns of partisanship and the economic vote.” *European Journal of Political Research* 50(3):365–394.

Krastev, Ivan. 2020. *Is It Tomorrow Yet? Paradoxes of the Pandemic*. Allan Lane.

Kunda, Ziva. 1990. “The case for motivated reasoning.” *Psychological Bulletin* 108(3):480–498.

Kuziemko, Ilyana, Ryan W Buell, Taly Reich and Michael I Norton. 2014. ““Last-place aversion”: Evidence and redistributive implications.” *The Quarterly Journal of Economics* 129(1):105–149.

Lewis-Beck, Michael S. and Mary Stegmaier. 2019. Economic Voting. In *The Oxford Handbook of Public Choice, Volume 1*, ed. Roger D. Congleton, Bernard Grofman and Stefan Voigt. Oxford University Press.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Annals of Applied Statistics* 7(1):295–318.

Lodge, Milton and Charles S Taber. 2000. Three Steps Toward a Theory of Motivated Political Reasoning. In *Elements of reason: Cognition, choice, and the bounds of rationality*, ed. Arthur Lupia, Mathew D McCubbins and Samuel L Popkin. Cambridge: Cambridge University Press pp. 183–213.

Malhotra, Neil and Alexander G. Kuo. 2008. "Attributing Blame: The Public's Response to Hurricane Katrina." *The Journal of Politics* 70(1):120–135.

Matějka, Filip and Guido Tabellini. 2020. "Electoral Competition with Rationally Inattentive Voters." *Journal of the European Economic Association* 19(3):1899–1935.

Olsen, Asmus Leth. 2017. "Compared to What? How Social and Historical Reference Points Affect Citizens' Performance Evaluations." *Journal of Public Administration Research and Theory* 27(4):562–580.

Park, Brandon Beomseob. 2019. "Compared to what? Media-guided reference points and relative economic voting." *Electoral Studies* 62.

Peterson, Erik, Sharad Goel and Shanto Iyengar. 2021. "Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign." *Political Science Research and Methods* 9(2):242–258.

Powell, G. Bingham, Jr. and Guy D. Whitten. 1993. "A Cross-National Analysis of Economic Voting: Taking Account of the Political Context." *American Journal of Political Science* 37(2):391–414.

Sacerdote, Bruce, Ranjan Sehgal and Molly Cook. 2020. Why Is All COVID-19 News Bad News? NBER Working Papers 28110 National Bureau of Economic Research.

Stroud, Natalie Jomini. 2008. "Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure." *Political Behavior* 30:341–366.

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

The Economist. 2020. "Only the world wars have rivalled covid-19 for news coverage." Last accessed: Jan 30 2023. `https://tinyurl.com/mfsx5877`.

Tilley, James and Sara B. Hobolt. 2011. "Is the Government to Blame? An Experimental Test of How Partisanship Shapes Perceptions of Performance and Responsibility." *The Journal of Politics* 73(2):316–330.

Wood, Thomas and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41(1):135–163.

Zhou, Rongrong and Dilip Soman. 2003. "Looking back: Exploring the psychology of queuing and the effect of the number of people behind." *Journal of Consumer Research* 29(4):517–530.

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

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.

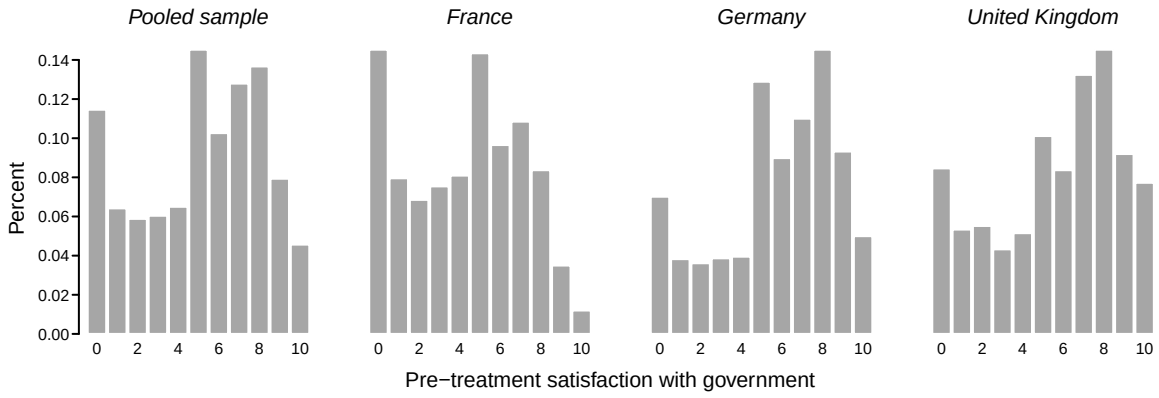

Figure A.1
Histograms of pre-treatment satisfaction with head of the executive

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.

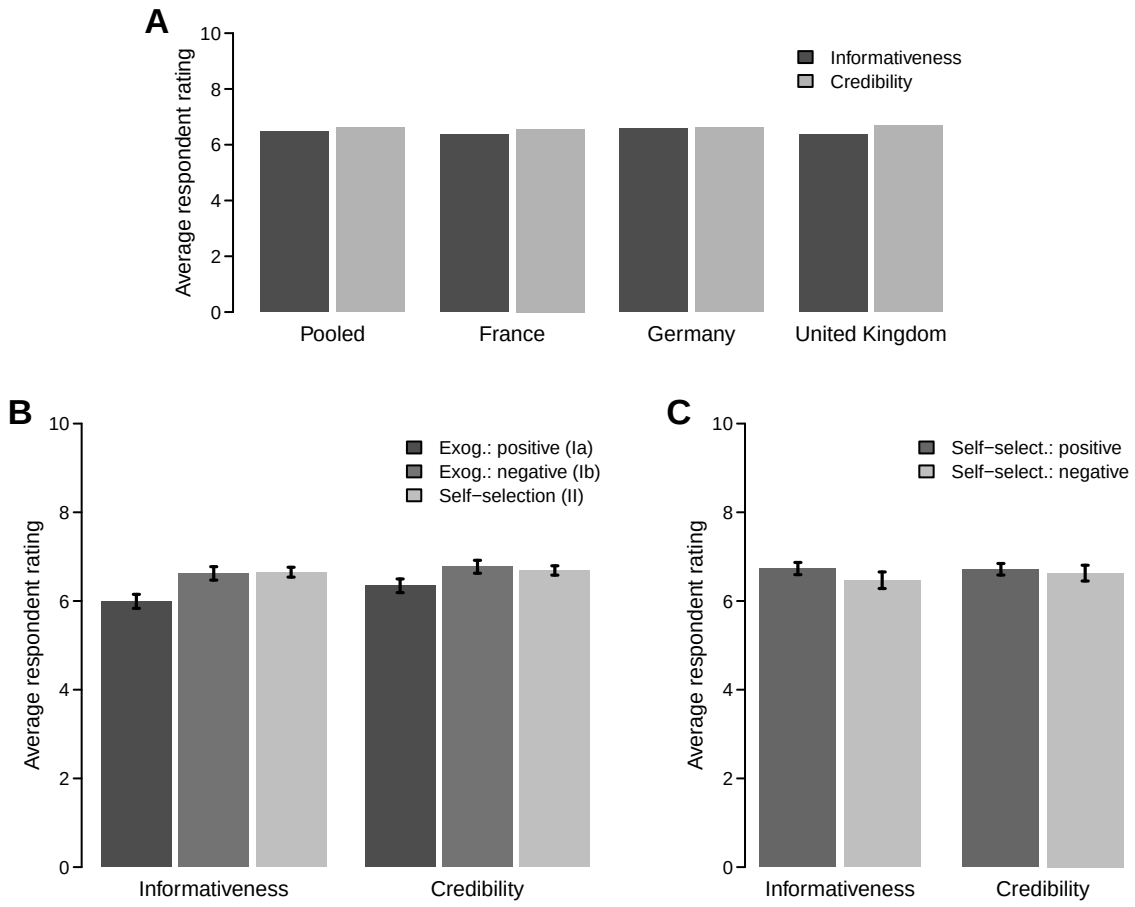

Figure A.2
Respondent evaluations of vignettes.

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.

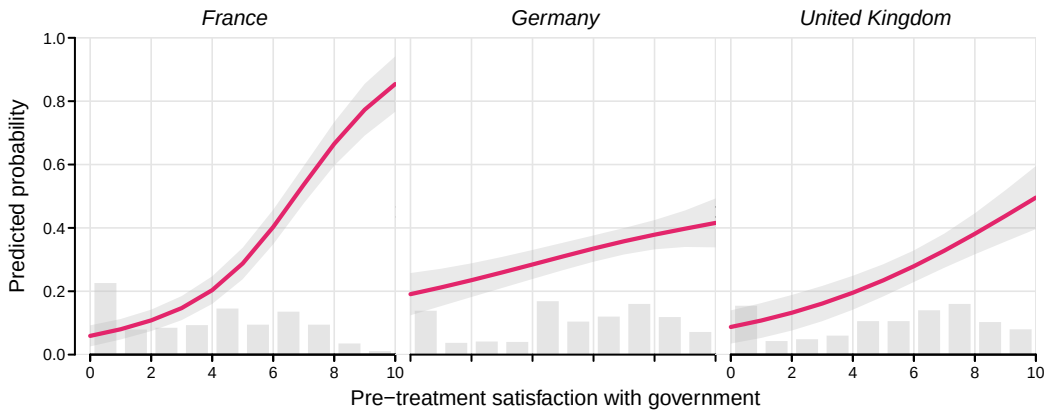

Figure A.3
Semiparametric model of the probability of positive benchmark choice as function of pre-treatment satisfaction with government

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).

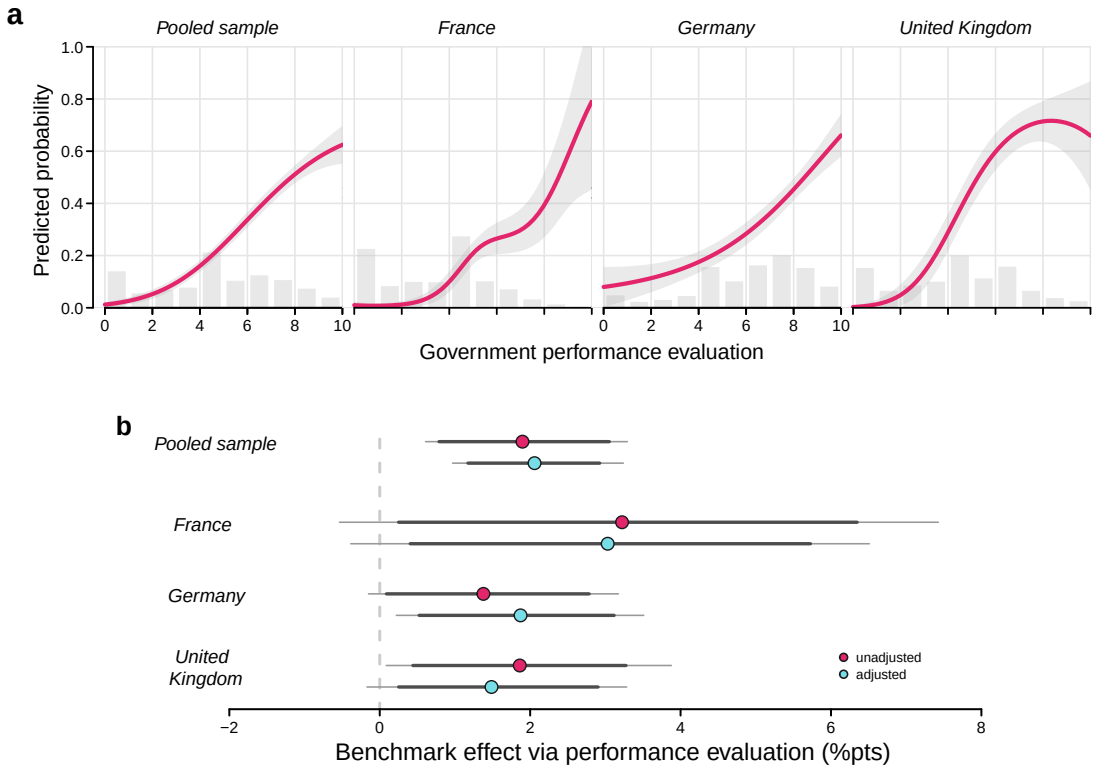

Figure A.4
Benchmarking information, performance evaluations, and vote choice

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.

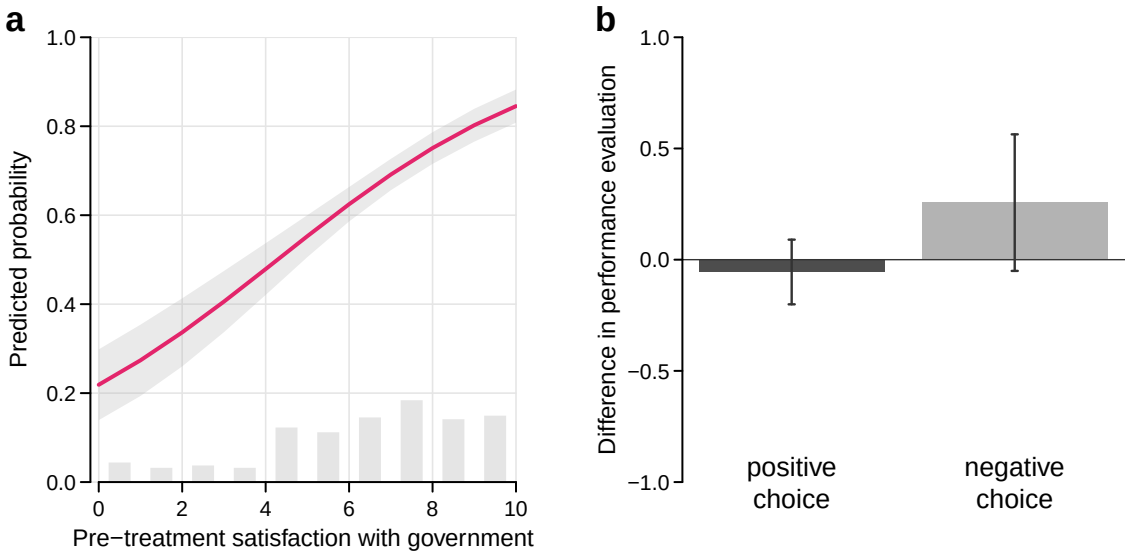

Figure A.5
Benchmark choice and information treatment effects in Austria

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.

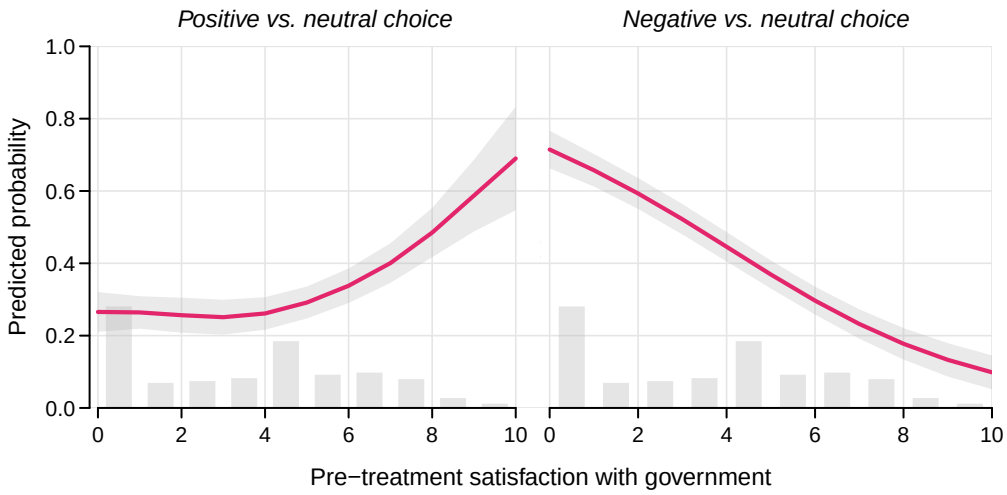

Figure A.6

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.