Article

Comparative Political Studies
2025, Vol. 58(4) 785–815

DOI: 10.1177/00104140241252095
journals.sagepub.com/home/cps


Worst of the Bunch: Visual Comparative Benchmarks Change Evaluations of Government Performance

William L. Allen[1,2] and Kristoffer Ahlstrom-Vij[3]

Abstract

Benchmarking theories argue voters use information about other countries' performances, usually on the economy and obtained through experience or media, to evaluate their own governments. Yet existing observational evidence is relatively fragile and struggles to distinguish how people become more knowledgeable. Using a pre-registered experiment, we showed UK respondents a chart displaying the UK's exceptionally high cumulative COVID-19 deaths either in isolation or alongside European countries with fewer deaths. Mimicking widely-circulated charts, this visual treatment enhances our study's external validity and tests the media-based channel for benchmarking. Aligned with pre-registered expectations, seeing the UK as "worst of the bunch" compared to UK-only data caused more negative government evaluations. Unexpectedly, partisanship did not moderate the information effects, while exploratory tests revealed the visuals generated

[1]University of Oxford, UK
[2]Nuffield College, UK
[3]Birkbeck, University of London, UK

Corresponding Author:
William L. Allen, Department of Politics and International Relations, University of Oxford, Manor Road Building, Manor Road, Oxford OX1 3UQ, UK.
Email: william.allen@politics.ox.ac.uk

Data Availability Statement included at the end of the article

more negative evaluations among respondents with high political trust. Our study shows international comparisons in visual forms can change domestic opinion, and on matters beyond strictly economic performance.

Keywords
benchmarking, COVID-19, government evaluation, political trust, UK, visualization

There has been long-standing interest in when and how voters ascribe responsibility for handling crises to different political actors, including their own governments (Hobolt & Tilley, 2014; Malhotra & Kuo, 2008). Although polarization along partisan lines also shapes patterns of blame attribution (Lyons & Jaeger, 2014) as well as other political outcomes (Iyengar et al., 2019), voters take information about incumbents' performance into account when forming attitudes (Carlin et al., 2021). Explaining how citizens use such information to hold their governments accountable is central for democratic functioning—particularly during crises (Bol et al., 2021; Jørgensen et al., 2021).

In the context of economic voting during downturns, benchmarking provides a potentially powerful mechanism by which citizens use comparative information to evaluate the economic performance of their own government against others, punishing incumbents' underperformance while rewarding overperformance compared to international peers (Aytaç, 2018b; Besley & Case, 1995; Powell & Whitten, 1993). Benchmarking has also been applied more widely to explain changes in support for European integration (De Vries, 2018; Martini & Walter, 2023). Yet the arguments for its presence attract several criticisms. Theoretically, its cognitively demanding assumptions strongly contrast with the common depiction of citizens as being generally uninformed about political and economic issues (e.g., Achen & Bartels, 2016; Lupia, 2015). Expecting voters to understand and act upon complicated economic data may set an unrealistically high bar. Empirically, much of the recent evidence for benchmarking relies on observational data involving choices of modeling and interpretation that subsequent re-analyses have shown to produce relatively fragile results, finding that alternative specifications using the same data produce null outcomes (Arel-Bundock et al., 2021).

While defenders of benchmarking have issued rebuttals to these charges, e.g., in Aytaç (2018a) and Kayser and Peress (2021), we cite the issues primarily as examples of how finding evidence of benchmarking in observational data can be difficult, which in turn suggests a need to also use experimental approaches (Healy & Malhotra, 2013). As it turns out, the

limited field and survey experimental evidence available suggests a greater plausibility for benchmarking in diverse contexts involving both hypothetical and realistic task environments (Aytaç, 2020; Bhandari et al., 2021; Hansen et al., 2015; Hart & Matthews, 2022). Therefore, the questions of whether citizens benchmark their governments' performances, how they do so, and under what circumstances are hardly settled.

In this paper, we contribute understanding about the benchmarking mechanism as a way by which voters form attitudes about their governments' performances—which has implications for subsequent behaviors and preferences. We do this by both building on the recent and fruitful focus on experimental approaches and using the case of COVID-19 which affords unique opportunities for exploring the micro-foundations of benchmarking. Specifically, we ask (1) whether information about a country's cumulative number of COVID-19 mortalities presented in isolation impacts how people perceive their government's handling of the crisis, and (2) whether these perceptions change when comparative information about other countries' own cumulative mortalities are included as well.

Drawing upon a pre-registered survey experiment ($N = 2917$ UK citizens) fielded in the UK through the wake of its first wave in July 2020, which was globally one of the worst in terms of mortality, we present two sets of findings. First, in line with our pre-registered expectation, people punished the incumbent government when its relatively poor standing among peers as the "worst of the bunch" in terms of cumulative deaths was made visible. Specifically, those who saw a graph that depicted the UK's COVID-19 cumulative mortality figures plotted alongside those of nine other European countries—all of which had similar leveling trajectories but fewer deaths overall than the UK at the time—expressed modestly more negative evaluations of the government's handling of the pandemic compared to those who saw the same graph without the comparative benchmarks. The direction and size of this main effect (about 0.1 of a standard deviation) corresponds with recent work on benchmarking during the pandemic (Becher et al., 2023 in France, Germany, and the UK; Martinez-Bravo & Sanz, 2023 in Spain; Shin & Park, 2022 in South Korea).

Second, we show that levels of political trust, but not partisanship, moderated respondents' evaluation of government performance given cumulative mortality data. Against our pre-registered expectation, we did not find heterogeneity with respect to partisanship in either of the conditions: effects among respondents supporting opposition parties were not more unfavorable compared to those supporting the incumbent Conservative party. However, in exploratory analysis separate from the pre-registered hypotheses, we found that high levels of political trust were associated with significantly *more* critical evaluations of government in both information conditions, compared to the effect at mean levels of trust. This suggests that political trust

may have acted as an expectation of a certain level of performance that was frustrated, rather than as a buffer of goodwill able to absorb perceived government failures.

Our design and case make several advances on existing work. First, since the spread of COVID-19 is a type of crisis on a scale previously unknown to most voters, it enables us to specifically assess whether and for whom providing explicitly *comparative* information matters for patterns of blame attribution. Economic benchmarking theories outline two possible means by which voters obtain this information: either from previous experience of economic turbulence (Lewis-Beck & Stegmaier, 2000) or from external sources including elites and media that "pre-benchmark" (Kayser & Peress, 2012) information in how they selectively report on the economy. By virtue of the unique global situation presented by the pandemic, our study provides a clearer test of whether conveying comparative information in modes mirroring those found in media has the expected benchmarking effect.

Second, throughout the pandemic to date, quantitative information has been particularly salient, arguably more so than in previous crises studied by political scientists such as the 2008 economic recession or natural disasters such as Hurricane Katrina. Moreover, this information has taken highly *visual* forms in media that clearly convey comparative dimensions (Falisse & McAteer, 2022; Financial Times, 2020; Roser et al., 2021) that elites have used for political purposes (Pentzold, Fechner, and Zuber 2021). For example, when the UK stopped displaying its national mortality figures against other European countries during its daily broadcasts from May 2020 onwards, the opposition party leader Sir Keir Starmer claimed "when we did not have the highest number in Europe, the slides were used for comparison purposes, and as soon as we hit that unenviable place, they have been dropped" (Hansard, 2020). This presents a clear political implication which we aim to test in this study: comparative benchmarks should matter for public evaluations depending on what they reveal about incumbents (Allen et al., 2024). For context, the UK continues to have the highest number of mortalities attributed to COVID-19 in Europe, standing at over 233,000 deaths as of December 14, 2023 (UK Government, 2023) placing it sixth globally (Our World in Data, 2024). Yet despite the obvious relevance of visual media for communicating information, few experimental studies have considered this format for benchmarking during COVID-19, with the exception of Martinez-Bravo and Sanz (2023) who use bar charts displaying proportions of COVID-19 contact tracers by region in Spain to indicate relative differences in this service. Crepaz and Arikan (2021) do convey Irish COVID-19 statistics in a visual dashboard, but their primary concern involves the effects of information disclosure rather than benchmarking in the sense we are pursuing. To be clear, given our design which did not include a text-only information condition, we

do not make claims about the distinctive effect of visuals compared to textual formats.

Third, our pre-registered experimental approach addresses concerns about how to attribute shifts in information to the expectations presented by benchmarking. Cross-sectional data—and decisions about how to separate signals of international performances from national ones—pose significant challenges to making causal claims. By contrast, we designed our treatments to tightly focus on the presence of comparative data presented in simple and visual forms that typify those used in digital media (Engebretsen and Kennedy 2020). Therefore, we contribute evidence of whether and how different forms of information impacts outcomes of significance to politics, and theoretically develop micro-foundational understanding about visual aspects of benchmarking.

Benchmarking and Evaluations of Incumbents During Crises

Our overall theoretical argument is that visual information revealing one country's relatively worse performance compared to its peers (a form of "benchmarking") causes citizens of that country to evaluate their government more negatively. We do so in three steps. First, we outline the assumptions of, and evidence bases for the benchmarking mechanism as a way of explaining incumbent evaluations—originally developed in contexts of retrospective economic voting. Second, we extend this logic into the visual realm and the domain of COVID-19. Third, we consider how partisanship and prior knowledge, which are factors shown to matter for patterns of blame attribution during crises, potentially place scope conditions on the effects of providing factual information on attitudes regarding government performance.

Benchmarking as an Explanation of Incumbent Evaluations

It is well-established that voters tend to punish incumbents during moments of crisis, either by blaming them for the problem or its perceived mishandling, rejecting them at the ballot box, or both (Ahlquist et al., 2020; Hobolt, 2015). But how and on what basis do they exert this accountability? A prominent debate considers whether evaluations of the government and attributions of responsibility are sensitive to information made available either through media or politicians themselves (Carlin et al., 2021; Damstra et al., 2020; Green et al., 2020; Malhotra & Kuo, 2008; Yagci & Oyvat, 2020). On the one hand, evidence suggests that citizens do take incumbents' prior performance into account when casting their votes, particularly on economic issues (Lewis-Beck & Stegmaier, 2000; Nadeau et al., 2012). On the other hand, there are good reasons to doubt the extent to which voters can and do behave this way.

Theoretically, such retrospective voting relies on people being able to accurately sense, sort, and act upon information—a requirement standing at odds with extant observations about widespread ignorance and misperceptions among electorates on a host of issues (Achen & Bartels, 2016; Lupia, 2015). In response, a third way argues that, while voters can take past performance into account as they evaluate their governments, this process faces several threats and scope conditions which may still result in mistakes (Duch & Stevenson, 2010; Healy & Malhotra, 2013).

Aligned with this middle option, we aim to develop the micro-foundations of benchmarking as a potential mechanism that links shifts in available information to evaluations. Benchmarking theories argue that, when voters learn about other peer countries' economic performances, they use this knowledge to form reference points from which they measure and either reward or punish their own government's performance (Aytaç, 2018b; Kayser & Peress, 2012; Powell & Whitten, 1993). Voters obtain this knowledge in two possible ways: either from previous experience of economic turbulence (Lewis-Beck & Stegmaier, 2000) or from external sources including elites and media that "pre-benchmark" (Kayser & Peress, 2012) information in how they selectively report on the economy. Therefore, voters should respond in predictable ways to information about their country's performance depending on the availability of relevant benchmarks and what those benchmarks reveal about the relative status of incumbents.

Yet this intuition, while appealing, has attracted criticism for reasons related to the broader debate on retrospective voting which benchmarking aims to explain. First, the evidence base for benchmarking largely comprises analyses of observational survey data that subsequent studies have shown to be fragile in the face of alternative model specifications and ways of separating national performance measures from global ones for which there no clear consensus (Arel-Bundock et al., 2021). Second, empirically distinguishing the predicted micro-foundations of prior experience from pre-benchmarking by elites is difficult to do using such cross-sectional data about economic issues. Third, regardless of which side one takes on these empirical questions, standard views on benchmarking still assume voters can accurately sense and evaluate abstract information, which may set an unreasonably high bar of competency.

Experimental approaches help to overcome the first of these concerns while probing the scope conditions under which benchmarking may occur (Healy & Malhotra, 2013). By exogenously providing information in controlled settings that is explicitly comparative, experiments can complement efforts to identify the causal effect of benchmarks. Indeed, recent work suggest that voters engage in behaviors consistent with benchmarking in both hypothetical and realistic situations (Aytaç, 2020; Bhandari et al., 2021; Hansen et al., 2015;

Hart & Matthews, 2022). Yet this does not address the second and third concerns.

Here is where our choice of issue domain and experimental design—detailed in the "Data and Methods" section—uniquely positions us to contribute to theories of benchmarking. Specifically, we use the context of COVID-19 which presents a valuable case for examining the media-based channel of benchmarking. This is because, as a global crisis whose scale and magnitude was previously unknown to electorates, respondents would not have had prior experience of pandemic-related events which could have also informed their evaluations.[1] Meanwhile, as we explain later, the timing of our experiment at the end of the first global wave in July 2020 but prior to widespread announcement of the successful Oxford/AstraZeneca vaccine means that we can be confident that our results reflect the opinions of UK-based respondents on how the government had handled a novel crisis that still did not have a clear end, as opposed to rewarding it for quickly producing a viable vaccine.

Relatedly, our *visual* treatment enhances the ecological and external validity of the experiment with respect to the media-based channel. Graphical visualizations that distill messages from complicated data are increasingly accessible by the public in various media (de Haan et al., 2018; Engebretsen and Kennedy 2020) and potentially impact political behaviors by changing the informational bases of voters' decisions (Allen, 2023; Young et al., 2018). As such, people are not assessing an entire complex dataset, but rather a simplified rendering that draws attention to key aspects for them (Amit-Danhi, 2022). This aligns with the expectation that media "pre-benchmark" (Kayser & Peress, 2012) information in the ways they selectively report on the economy. Moreover, the comparative line graphs showing COVID-19 mortalities which we replicate in our experiment were particularly visible and widely referenced, including in daily national UK press conferences during the first wave (Allen et al., 2024), and usually accompanied by commands to "flatten the curve" (Pentzold, Fechner, and Zuber 2021). As countries' case records extended, these visuals served as sources of information that citizens could draw upon to evaluate how well their government was handling the pandemic compared to other countries.[2]

To date, research suggests a strong plausibility for benchmarking occurring during the first wave of the pandemic. For instance, evidence drawn from European surveys about incumbent support during the early stages of the pandemic suggests that citizens were already paying attention to events happening in other countries—such as Italy's lockdown in March 2020—even as their own governments' course of action was unclear (De Vries et al., 2021). Meanwhile, observational survey data from South Korea in April 2020 demonstrates how citizens holding more favorable views of the government's *relative* COVID-19 performance were more likely to express higher

presidential approval ratings (Shin & Park, 2022). This pattern coincides with results from cross-national observational European data (Bol et al., 2021) and experimental data collected across France, Germany, and the UK (Becher et al., 2023) around the same time.

Factors that Moderate Information Processing: Knowledge and Partisanship

However, there are two important factors that potentially moderate how citizens process information, coming from the distinct yet related literature examining how voters attribute blame to their governments during crises. First, people holding different levels of political knowledge engage in different patterns of blame attribution, with less knowledgeable individuals being more likely to blame a government currently in office—even if it had little to do with a given crises—because it is more immediately familiar and proximate to the problem at hand (Gomez & Wilson, 2008; Luskin, 1987). Conversely, more knowledgeable individuals tend to divide responsibility among several recipients, especially in complex crises with multiple actors.

Second, partisanship can lead to biases in blame attribution: partisans opposite to the incumbent are more likely to attribute fault to the government and negatively evaluate its performance (Hellwig & Coffey, 2011; Tilley & Hobolt, 2011). This has been shown to also occur in moments of economic crisis (Bisgaard, 2015): while citizens may agree on factual statements about a situation, they still resort to partisan reasoning as they form opinions and assign blame (De Vries et al., 2018; Lyons & Jaeger, 2014). However, whether this reasoning held during the early period of the COVID-19 pandemic in which we conducted our study is unclear. On the one hand, during the initial wave of the pandemic, research from the US context suggested that partisans viewed the pandemic through very different lenses: in March 2020, Democrats were much more likely to adopt behaviors in line with public health guidance compared to Republicans, as well as express greater concern about the pandemic's consequences for themselves and the country (Gadarian et al., 2020). Meanwhile, longitudinal survey evidence from European countries including the UK shows how citizens who voted for the incumbent government in the previous national election were more likely to support their government's handling of the crisis during 2020 (Jørgensen et al., 2021). On the other hand, experimental evidence from the US around the same time as our study shows that emphasizing then-President Trump's handling of the pandemic did not impact beliefs, attitudes, or behavioral intentions (Spälti et al., 2021).

Hypotheses: Visual Comparative Benchmarks Change How Voters Evaluate Government Performance

Given this background, our study considers whether such visual representations of longitudinal information about cumulative COVID-19 mortalities impacts evaluations of incumbents even after accounting for a variety of factors including partisan leanings and self-perceived numeracy.[3] Although leaders often initially benefit from "rally-round-the-flag" effects in the face of shocks and crises (Mueller, 1970), voters evaluations generally turn negative as time goes on. This has also been in the case with COVID-19: multiple presidents and prime ministers saw large increases in their public support immediately after the World Health Organization declared COVID-19 to be a pandemic and governments imposed restrictions such as lockdowns (Bol et al., 2021; Jennings, 2020) that have since worn off (e.g., Johansson et al., 2021).

Within this context, our pre-registered hypothesis was that high mortality figures—more than 40,000 in the UK at the time of our experiment—would be perceived as a straightforward sign of the government's failure in handling the crisis.[4] Moreover, if benchmarking were occurring, seeing information confirming a country's status as the worst-performing among peers on this metric should cause *even more* negative evaluations.

H1a. (*high mortalities*) The presence of mortality data will cause participants in either of the two treatment groups to express worse perceptions of government performance, compared to the control group seeing no image.

H1b. (*benchmarking*) Those seeing comparative mortality data showing the UK's exceptional status among other European countries will express worse perceptions of government performance, compared to those who see the UK-only data.

However, we acknowledge the potential for partisan motivated reasoning to bias how some respondents will interpret the information (Bisgaard, 2015), although recent work questions the likelihood of heterogeneous effects (Coppock, 2022). Respondents whose partisanship matches the government might be more likely to discount the negative information, and consequently be less likely to attribute blame for the situation to the government's actions (Jørgensen et al., 2021). As such, we expected these effects would interact with partisanship:

H2a. (*partisanship*) In either treatment group, partisanship will moderate the effect, with a difference in effect towards more favorable perceptions of government performance among people whose political identification

matches that of the incumbent government (i.e., the Conservative party), compared to people supporting opposition parties (i.e., Labour, Liberal Democrat, Green Party, Scottish National Party, or Plaid Cymru) or no party.

We also implemented an exploratory interaction hypothesis after data collection, following from work that emerged at the beginning of the pandemic and during our period of analysis on the role of political trust in shaping attitudes and preferences (Bavel et al., 2020; Devine et al., 2021, 2023). Considering this work, we examined whether respondents' levels of political trust mattered for the information treatment effects. Prior scholarship shows how, in ordinary times, citizens' levels of trust in government institutions matter for their perceptions of performance: less trust leads to less positive evaluations (Chanley et al., 2000). Global perceptions during the initial wave of COVID-19 also displayed a positive association between levels of trust and evaluations of how governments were handling the pandemic (Lazarus et al., 2020; Narlikar & Sottilotta, 2021). Moreover, experimentally disclosing information about the dynamics of the early stages of the pandemic—including comparative information displayed in visual dashboards that typified publicly available outputs at the time—appeared to create a feedback loop whereby people holding already low levels of trust expressed even more negative views of authorities (Crepaz & Arikan, 2021).

Therefore, we expected that respondents holding lower levels of trust in government would be more affected by information conveying high mortality levels. Since high-trusters would presumably be pre-disposed to believe that government generally has their best interests at heart, this pre-dispositioning would serve to cushion to blow of the suggestion that government is not performing at the level that it ought. We operationalized high levels of trust as one standard deviation above the mean on our trust scale (see Data and Methods) and took the contrast to be the mean level of trust.

H2b. (*political trust*) In either treatment group, level of trust in government will moderate the effect, with a difference in effect towards more positive perceptions of government performance among those exhibiting high trust in government, compared to the effect among those at lower levels of trust (exploratory).

Data and Methods

Experimental Design and Treatments

We designed and pre-registered a between-subjects survey experiment in which we randomly assigned respondents to one of three conditions.[5] In the *Control*

condition, respondents answered standard sociodemographic and political questions, as well as a series of questions measuring personal contact with COVID-19, knowledge of COVID-19 (World Health Organization n.d), trust in government (using a battery validated by Devine et al., 2020), and self-perceived numeracy (using a version of the Subjective Numeracy Test validated by McNaughton et al., 2015).[6] We included self-perceived numeracy because this might condition how people engage with statistical data, and constructed our scale using mean responses in line with established practices. Meanwhile, we constructed scales for personal contact with and knowledge of COVID-19 using Item Response Theory (IRT) and the R package *mirt* (Chalmers, 2012) since these also potentially matter for government evaluations as we describe in the theory section (Jørgensen et al., 2021).[7]

Meanwhile, participants in the *UK-only* condition saw an image depicting cumulative mortality data for the UK since 50 cumulative deaths were recorded in the country (upper panel of Figure 1).[8] We plotted this on a linear scale because we thought it would be more accessible to respondents than logarithmic scales, and experimental evidence suggests scaling choice does not impact our outcome of interest (Sevi et al., 2020). Finally, in the *Comparative* condition, respondents saw an identical image (lower panel of Figure 1) containing three key differences: (1) cumulative deaths plotted for nine additional European countries, (2) an additional subtitle saying, "Top ten European countries," and (3) an additional footnote explaining that "country data is aligned by stage of the outbreak." These differences drew attention to the comparative nature of the image, as well as the UK's unique position as having the most COVID-19 mortalities relative to other European countries as of July 15, 2020.

Sampling Procedures

We recruited 3018 participants between July 15–17, 2020 using Prolific.ac, which holds an established UK panel that displays advantages over other options such as MTurk in terms of respondent diversity, pre-screening, and data quality (Peer et al., 2017, 2022). Fieldwork concluded before the widespread public announcement of the successful Oxford/AstraZeneca vaccine trial on July 20, 2020 (Folegatti et al., 2020). This removes concerns about our results being influenced by favorable news about the positive results of the trial, which was prominently funded and promoted by the government. We aimed to achieve a sample that was representative of the UK population in terms of gender and non-White ethnic minority categories. Ethnicity was important because widely-publicized evidence shows Black, Asian, and other ethnic minority groups are more likely to die from COVID-19 than White British or White Irish patients in England (Aldridge et al., 2020). Comparing our sample to the June 2020 wave of the British Election Study (BES), commonly viewed as a gold-standard survey dataset for the UK electorate, we observe how our sample is somewhat younger, more educated, more likely to

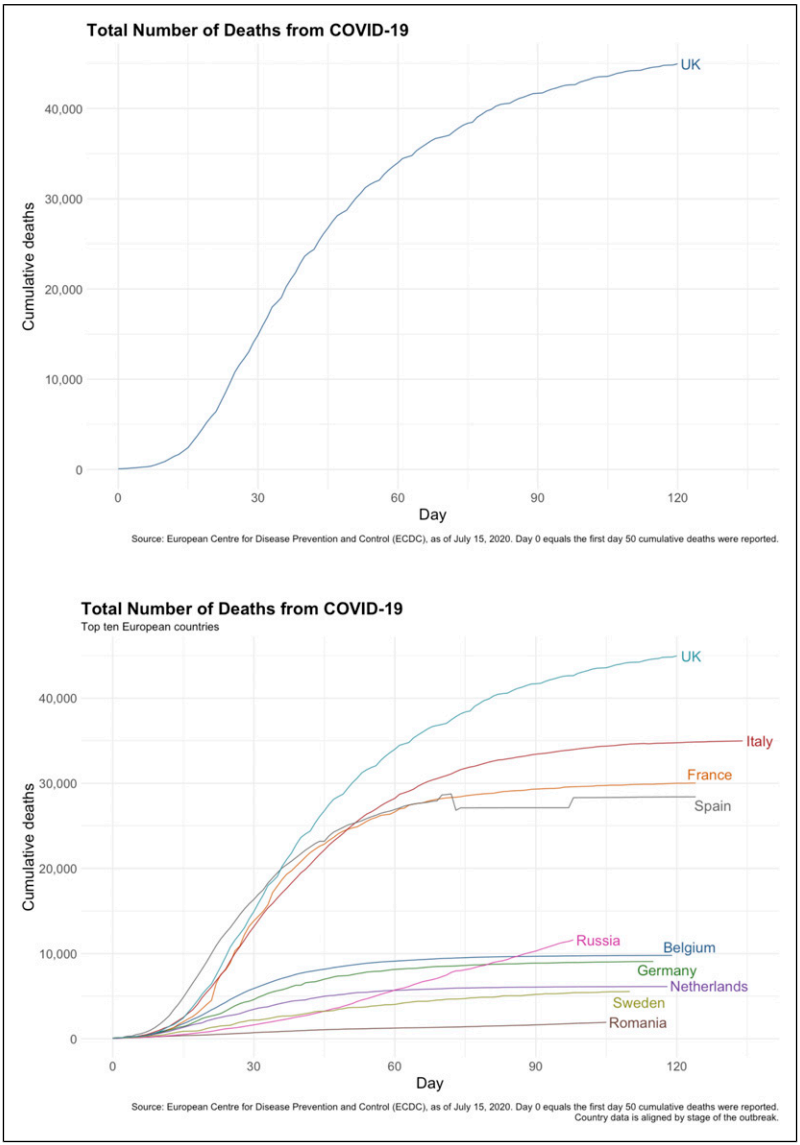

Figure 1. Treatment images used in the UK-only (top) and Comparative (bottom) experimental conditions.

support the Labour Party, and more likely to have voted "Remain" in the 2016 EU referendum than reported in the BES sample (see Table 1 in the supporting information). Nevertheless, as reported in the supporting information and pre-registered plan, we have adjusted for these and other covariates, all of which were measured pre-treatment. We set up 12 calls for participants on Prolific, with each call corresponding to one cell in the cross-tabulated survey targets and opening at different times which were for all intents and purposes random. Participants responding to one call were not notified of the others, so participants could only enter through one of the links. Differences among our three groups along gender and ethnicity were minimal.[9]

Following our pre-registration protocol, we dropped respondents who spent fewer than 5 seconds on either treatment image page (0 respondents), which we reasoned was a minimal amount of time to look at the graph. We also disqualified any respondents who managed to take the survey twice (3 respondents comprising 6 observations), and any who failed to register their age (3 respondents)—a disqualification we had not pre-registered, but one we deemed more reasonable than attempting to impute these participants' ages. At the pre-registration stage, we decided to remove respondents who failed a manipulation check at the end of the survey: we asked individuals to indicate whether one or many countries were displayed in the image they had seen earlier, and dropped those answering differently from the treatment group to which they were assigned (i.e., anyone answering "many" in the *UK-only* group or "one" in the *Comparative* group, 92 respondents in total). However, acknowledging that dropping respondents who fail a post-treatment manipulation check can introduce bias (Montgomery et al., 2018), we also conducted the analysis using the full sample which deviated from our pre-registered plan. Our results remain the same under those conditions, except for two instances where non-significant estimates become significant (see footnotes 13 and 17). Our final sample comprised 2917 respondents: 1001 in the *Control*, 921 in the *UK-only* group, and 995 in the *Comparative* group.

Outcome Variable: Government Evaluation

We report treatment effects for the outcome of *perception of government performance*, which is a continuous IRT model scale analyzed by way of a linear model.[10] This outcome is measured by agreement with six statements on a 1–5 scale, 1 being "strongly disagree" and 5 being "strongly agree." These items, taken from an Ipsos (2020) poll fielded earlier in the pandemic, include statements such as "The UK government has done a good job of protecting UK residents through its response to the coronavirus" and "The UK government's plan has adapted well to the changing scientific information and situation." Note that, since this scale had not been pre-validated through

previous studies, we followed our pre-analysis plan and only used the subset of items that made for a valid IRT model (see Supporting Information).

Results

Main Effect: Do Mortality Statistics Change Government Evaluations?

Our first step involves asking whether the information treatments had effects on perceptions of how the government was handling the pandemic. Following our pre-registration plan, we fitted a linear model using *lm* in R's *stats* package (R Core Team, 2018) to estimate the (main) effect of being in one of the treatment groups, compared to the control, while holding constant all covariates.[11] Figure 2 (top two lines) shows our main effect on government performance.

The information contained in the treatments impacted respondents' evaluations of the government, but not entirely in the ways we had hypothesized.[12] Neither treatment group expressed more negative evaluations that reached our pre-registered significance level ($p < .05$) compared to the

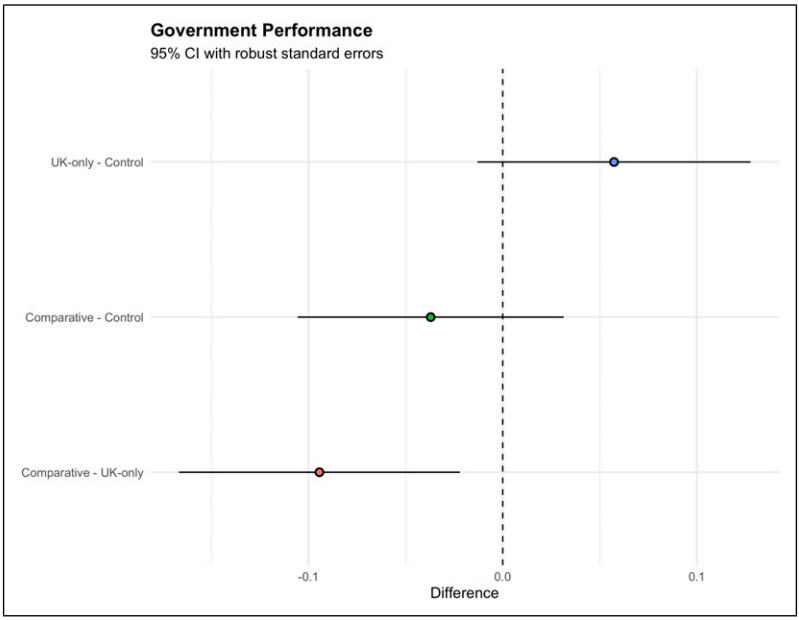

Figure 2. Treatment effects plotted with 95% Wald confidence intervals using robust standard errors and Holm correction for multiple comparisons.

control group, which does not lend support for H1a (*high mortalities*). Rather, seeing the UK's mortality data in isolation may have led respondents to evaluate the government *more positively* than those in the control condition, although this estimate also did not reach our pre-registered level of significance when restricted to those who did not fail the manipulation check (p = .057).[13] Our logic at the pre-registration stage was that the high cumulative numbers of COVID-19 deaths in the UK (i.e., gross levels) would be the most salient feature of the UK-only graph, and therefore cause more negative evaluations of the government in charge. However, it appears that the *rate of growth* might have been more salient for respondents in the UK-only condition, who as a result perhaps rewarded the government for its apparent success at "flattening the curve." Reflecting on this, we acknowledge how public health officials and media had widely promoted this objective in the early part of the pandemic (Roberts, 2020).

Comparing the two information conditions provides more direct evidence that benchmarking is occurring, and that the rate of growth was probably the more salient feature in the UK-only image (bottom line in Figure 2). The difference between the UK-only and Comparative treatment conditions is statistically significant and follows the direction hypothesized in H1b (*benchmarking*): respondents seeing the UK's cumulative mortalities set against other European countries—all of which were lower for most of the period covered—expressed more negative evaluations of the UK government than those seeing the UK figures in isolation by 0.09 standard deviations (SE (robust) = 0.031, p = .007).[14]

We argue this is evidence of benchmarking because the only meaningful difference between these treatments was the presence of lines indicating other countries' mortality data that were all lower than the UK in terms of their levels but practically identical in terms of their slopes (rates of change). In this instance, comparative data appears to cause respondents to punish the government, although to a modest degree which seems sensible given expectations about the ability of a single graphic to substantially change attitudes in an information-rich environment.[15] Given that most of the other countries *also* had flattened their own mortality curves, this marker of success—rewarded in the absence of comparisons by the first treatment group—no longer seems distinctive and positive. Instead, information signaling peers' performance serves as a reference point for critical evaluations.

While we have argued that our visual treatment design and the pandemic context of the study provide a uniquely relevant test for the media-based channel of benchmarking, knowledge of and experience with COVID-19 during the first wave of 2020 prior to our experiment may still have been relevant for respondents' evaluations of the government (see footnote 1). To check for this possibility, we ran additional models in an exploratory (i.e. not pre-registered) manner which interacted each treatment with respondents' IRT

scores on the pre-treatment COVID-19 contact and knowledge batteries (Tables 7 and 8 respectively in the Supporting Information). In both instances, we do not find significant interaction effects: respondents who had more personal contact with the pandemic, or knew more about COVID-19, did not express different evaluations compared to those with less contact or knowledge respectively after seeing either treatment.

The Roles of Political Trust and Partisanship in Evaluating Government Performance

As suggested by prior theories of blame attribution, our main information treatment effects on government evaluation may interact with different levels of trust in government and partisanship. In terms of partisanship, we hypothesized at the pre-registration stage that there would be a significant difference in effect on perceptions of government performance among partisan subgroups, with the effect among supporters of the incumbent Conservative party being towards more positive evaluations compared to the effect among supporters of opposition parties or no party. Figure 3 shows how none of the differences were significant.[16] Therefore, we reject H2a (*partisanship*). This null finding corroborates experimental evidence of attitudes held by heterogeneous groups moving in the same direction on account of exogenously-varied information treatments (Becher et al., 2023; Coppock, 2022).

Next, contrary to what we had hypothesized in an exploratory manner, the moderating effect among high-trust participants—i.e., those who were one

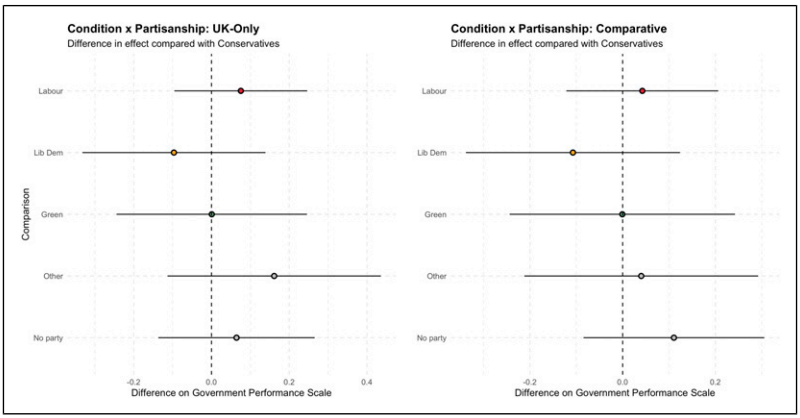

Figure 3. Difference in treatment effect on government performance by partisanship and treatment (UK-Only and Comparative) compared with Conservatives, with 95% Wald confidence intervals using robust standard errors.

standard deviation above the mean on the political trust scale—in both treatment conditions was towards *less* favorable views of government performance, compared to those at the mean level of trust. Consequently, we reject H2b (*political trust*). Specifically, while perceptions of government performance on behalf of participants exhibiting mean levels of trust in the two conditions were not significantly different from the control ($B = 0.056$, SE (robust) = 0.030, p-value = .063 for UK-only; $B = -0.037$, SE (robust) = 0.029, p-value = .209 for Comparative),[17] there was a significant *difference* in effect introduced by the interaction term ($B = -0.108$, SE (robust) = 0.031, p-value = .001 for UK-only; and $B = -0.078$, SE (robust) = 0.031, p-value = .013 for Comparative) when comparing those at the mean level of trust to those one standard deviation above that mean.[18] This interaction can also be seen from the distinct slopes in Figure 4.

This is noteworthy because it questions the assumption that high levels of trust serve as a buffer of goodwill that protects people's sense that government is performing at the level that it ought. Had that been the case, we should have seen a *positive* coefficient value for the trust interaction, making for a steeper slope for the treatment slopes than the control slope. Instead, the perceptions of respondents exhibiting high levels of trust seem not so much protected by the presence of trust as making it vulnerable to disappointment. Returning to prior research into the causes of political trust lends some support to this: if citizens' trust in government institutions is informed by perceptions of good performance on the part of those same institutions (Mishler & Rose, 2001), then information drawing attention to those perceptions being violated is likely to have negative consequences. By contrast, low trust individuals perhaps do not assume that government will perform well to begin with, are less vulnerable to disappointment as a result, and—as in the UK-only condition—therefore might have been unexpectedly surprised that the government was successfully flattening the curve.

Discussion

At the outset of the paper, we highlighted the example of an opposition politician in the UK pointing out apparent selectivity on the part of the incumbent government when it came to reporting COVID-19 mortalities. The implication was clear: comparison is politically useful if it shows the government in a positive light compared to its neighbors. Indeed, several countries' governments invoked their comparative performance on several COVID-19 metrics over time, notably New Zealand under Prime Minister Jacinda Ardern who received widespread public praise for her initial handling of the crisis (Gilray, 2021). This is consistent with existing theories of benchmarking as applied in the domain of economics, which suggest that voters are sensitive to comparative economic information because it provides

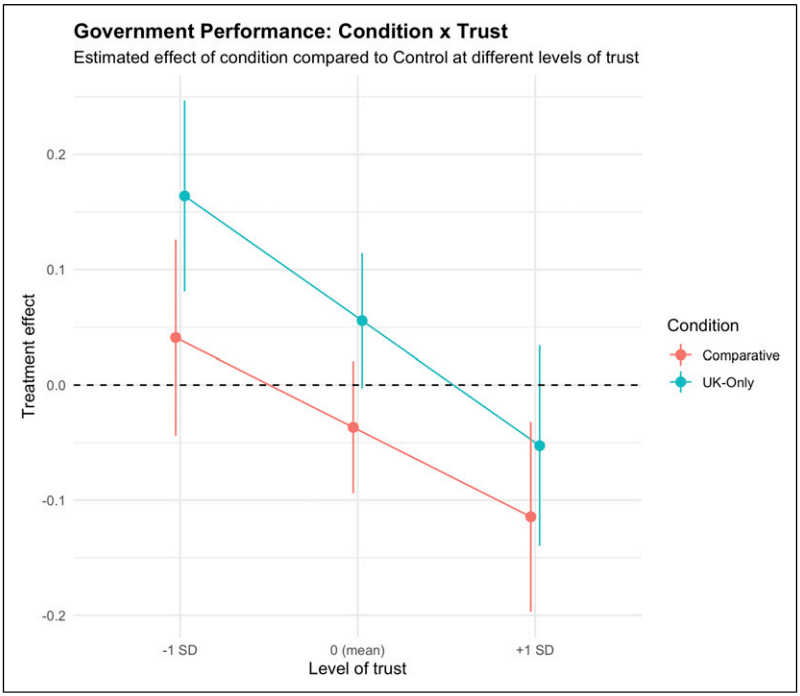

Figure 4. The effect of each condition (UK-Only and Comparative) on perceptions of government performance compared to Control at different levels of trust (0 corresponds to the estimated mean level of trust in the population). 95% confidence intervals calculated with robust standard errors.

a reference point that contextualizes how their own governments are performing, particularly during downturns (Besley & Case, 1995; Kayser & Peress, 2012). However, there are important theoretical and empirical challenges to this argument, notably the consequences of complex modeling choices (Arel-Bundock et al., 2021). Moreover, we know relatively little about the micro-foundational role of media in pre-benchmarking comparative information, since distinguishing this from prior knowledge or experience is difficult using cross-sectional data about economic issues.

We have addressed these problems by way of a pre-registered experiment using the novel context of visually communicating public health statistics during COVID-19, a different kind of crisis compared to those previously studied by political scientists interested in benchmarking. Specifically, using the case of the UK—a country that has had exceptionally high numbers of COVID-19 deaths among European countries—we expected that visualizing these alarming statistics at the time of our experiment in a simple way would

cause respondents to express more negative evaluations of the government's handling of the crisis. This expectation was partly borne out by our results. First, when respondents lacked international data that could have served as benchmarks, respondents seemingly focused on the rate of growth rather than gross levels. As a result, when compared to participants in the Comparative condition, respondents *rewarded* the incumbent administration for apparently flattening the country's mortality curve. However, when other European countries' *own* flattening trajectories were visible in the comparative condition, the UK's status as "worst of the bunch" in terms of cumulative mortalities likely became more salient. Consequently, and in line with expectations of benchmarking, respondents punished the incumbent government in the form of modestly lower performance ratings compared to those seeing UK-only data, on the order of about 0.1 of a standard deviation on our IRT scale measuring government approval. This modest effect size—which matches what Becher et al. (2023) find in their pooled experimental results across France, Germany and the UK—seems reasonable given expectations about the power of a single graphic in an information-rich environment to change citizens' attitudes. Therefore, as to whether it was advisable for the UK government to drop comparisons of its national mortality figures with other European countries from its daily press briefings, this would seem to have been a well-calculated move from a political point of view.

More fundamentally, our study makes two theoretical contributions: on the one hand, to understanding the scope conditions of when (and which) voters attribute blame to their governments during crises; and on the other hand, to understanding about benchmarking as a specific means by which voters respond to information as they evaluate their governments' performances. First, focusing on blame attribution, we find that levels of trust in government likely moderates *who* engages with information and with what effects. We found a significant interaction between our treatments and levels of trust on perceptions of government performance, with the effect among high-trust participants in both treatment conditions being towards *less* favorable views of government performance, compared to the effect at mean levels of trust. This suggests that public trust is not necessarily a reservoir of goodwill that governments can tap into in tough times. Rather, any trust gained by perceptions of unexpectedly good performance might just be raising the bar and in so doing create expectations that can easily be frustrated when government no longer performs at the level that they have led people to expect.

Notably, our results with respect to trust differ from other experimental work conducted slightly earlier in Ireland, which found that respondents already holding low levels of trust in government were more likely to ascribe even *less* trust in government after seeing detailed information about COVID-19—including how Ireland's cases compared to other countries (Crepaz & Arikan, 2021). One possible explanation lies in the complexity of information

displayed in each study: while Crepaz and Arikan (2021) find that pre-treatment levels of trust significantly interact with the effects of their "high information" treatment which provided nearly a dozen data series in a dashboard format, they do not find a similar interaction effect for their "low information" treatment which only showed Irish data about tests, cases, patients in intensive care, and mortalities. If concerns about greater information disclosure actually leading to less public trust are warranted (Worthy, 2010), then it is possible that exposing respondents to large amounts of data—particularly about anxiety-inducing and previously unfamiliar topics such as hospitalization rates—might trigger suspicion and have further consequences for behaviors. This is partially seen in the context of COVID-19: survey experimental evidence among US and Danish citizens showed how transparent yet negative communication about the pandemic lowered respondents' acceptance of vaccines, but actually increased their levels of trust in health authorities (Petersen et al., 2021).

Another possible explanation lies in the extent to which respondents associated the treatments with either the political or administrative arms of the state. Crepaz and Arikan (2021) intentionally based their treatment dashboards on an existing website managed by the Irish Department of Health. By contrast, our treatment only referenced the non-partisan European Centre for Disease Prevention and Control as the source of the data, and did not use any established branding or formatting that could be connected to the UK government's COVID-19 messaging (see Figure 1). Public opinion scholarship demonstrates how audiences turn to messengers they perceive to be credible for guidance on which forms of information they should believe (e.g., Druckman, 2001). Therefore, showing low-trust respondents a message imitating a website backed by a government they already do not believe to be credible might indeed lead to more negative evaluations, whereas a message lacking clear branding—or indicating a non-partisan messenger—could have no or even opposite effects. As such, we think these sets of findings may be consonant with each other, while also warranting further investigation to account for treatment complexity and perceived messenger partisanship.

Second, our findings also add to theorization about benchmarking as a means of explaining how voters hold their governments to account using external information on a host of politically relevant issues, while also lending urgency for taking international comparison seriously as a rhetorical and political tool (Broome et al., 2018) —one that is increasingly applied to matters extending beyond economic performance as other empirical studies have shown (De Vries, 2018; Martini & Walter, 2023). Actors in media and policymaking can readily access comparative datasets relating to other areas of governance such as international development (Bandola-Gill et al., 2022) and security (Baele et al., 2017). This raises the stakes for understanding whether accessing these resources matter for how citizens evaluate their

governments' performances—and ultimately processes of selecting and sanctioning politicians by way of voting behaviors which flow from these evaluations. On this front, more evidence on *which* benchmarks citizens pay attention towards—and whether these choices are related to prior beliefs—is needed (Becher et al., 2023).

We have also demonstrated the importance of *comparative* and *visual* representations of data as specific types of information to which voters can respond. While much work on information effects using observational data has shown how becoming more informed about political and economic issues is associated with shifts in attitudes and voting behaviors (Ahlstrom-Vij, 2021; Carlin et al., 2021), these findings are largely based on implicitly *national* and text-based forms of information. By contrast, our study highlights how citizens' information environments are increasingly international while also involving new media that aim to make complex information more understandable in a form of pre-benchmarking (Bekkers & Moody, 2014; Engebretsen and Kennedy 2020). Future work should consider how and to what extent these kinds of visual media impact political perceptions, attitudes, and preferences.

At their broadest level, our results show how factual information matters for government evaluations *despite* several unique features of the COVID-19 pandemic compared to previous economic and humanitarian crises studied by political scientists such as 9/11 and Hurricane Katrina: a global and acutely-felt event that generated extraordinary restrictions on mobility and daily life; an unprecedented salience of statistical information conveyed by mass media and political rhetoric; and an explicitly comparative dimension to that information which was used to support conclusions about countries' relative performances. As such, future work will need to trace how and in what ways the political behavioral impacts of this crisis differ from established theories about blame attribution.

Acknowledgments

Previous versions of this paper were presented at the 2020 virtual conference "Political Trust in Crises" (University of Southampton, October 22–23), the 2021 virtual European Political Science Association Annual Meeting (June 24–25), and the 2023 International Communication Association Annual Conference (Toronto, May 25–29). We would like to thank the discussants and participants at these events for their valuable feedback.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for the survey experiment was provided by Birkbeck, University of London and Magdalen College, University of Oxford. William Allen acknowledges support from the British Academy (grant number PF21\210066).

ORCID iD

William L. Allen https://orcid.org/0000-0003-3185-1468

Data Availability Statement

Data and replication materials are available in the Harvard Dataverse at https://doi.org/10.7910/DVN/VMZIPF (Allen and Ahlstrom-Vij, 2024).

Supplemental Material

Supplemental material for this article is available online.

Notes

1. We acknowledge respondents would have had information about how governments had responded to the crisis in the short-term, i.e. during the first wave in 2020, which may have been relevant for their evaluations. As we explain in the methods and empirical results, we account for this possibility by way of pre-treatment question batteries about respondents' level of personal experience with, and knowledge of COVID-19.
2. As the World Health Organization correctly warns, making strong comparisons across countries is inadvisable due to different ways of counting and categorizing COVID-19 cases and deaths. While we acknowledge the difficulties in accurately interpreting cross-national statistics in this context, our study is premised on the undeniable observation that citizens and politicians nevertheless *did* attempt to draw conclusions from comparative data.
3. Unless otherwise labeled as "exploratory," all hypotheses were pre-registered.
4. In the full study, we also included pre-registered hypotheses relating to public health-related outcomes: concern about the pandemic, support for restrictions, and anticipated behaviors. While we focus on the government evaluation outcome because it directly tests the benchmarking mechanism we are interested in, we provide results for these other outcomes in the Supporting Information. In short, the presence of comparative data as reference points did not significantly change any of these outcomes in our experiment.
5. Replication materials and code can be found in Allen and Ahlstrom-Vij (2024). The pre-registration is at: https://osf.io/tg6nz/?view_only=71fe2b6a5ef74bd098935cddf25fe4a5.
6. The full range of questions used appears in the Supporting Information.

7. IRT models measure latent traits assumed to fall on a continuous scale that has a mean of zero and standard deviation of 1, meaning that while values ascribed to a respondent have no intrinsic meaning, they can nevertheless be interpreted relative to an estimated population mean (details and diagnostics appear in the Supporting Information).
8. The mortality data came from the European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. Plots were created using R's *ggplot2* package (Wickham, 2016).
9. See Tables 1 and 3 in the Supporting Information document for descriptive statistics. We spread out recruitment because the calls targeting the same demographic groups (but for different treatment conditions) needed to be launched and completed serially in order to exclude anyone who had been invited to a previous call from being re-invited, and thereby ending up in two treatment groups. We have no reason to believe this biased our results.
10. As described in footnote 4, we also constructed three other outcome variables using IRT modeling approach. See Supporting Information for full details on the IRT models, where all statement prompts can also be found.
11. Model details and diagnostic for all main effects can be found in Table 4 in the Supporting Information. Table 10 includes model details for all main effect models fitted without covariates. Although noting the very poor fit of these models, we observe the relative effects between the two conditions remain the same as with covariates included: a higher coefficient value in the UK-only condition compared to the Comparative condition in all conditions, except for levels of concern, where the effect is about the same.
12. We used the *glht* (general linear hypothesis) function from R's *multcomp* package (Hothorn et al., 2020) to perform pairwise comparisons with Tukey contrasts and the Holm method for p-value correction. We also used robust standard errors as calculated with the *sandwich* package (Zeileis, 2004).
13. When no participants were dropped (see reasoning in "Sampling procedures" above), this estimate turned significant: $B = 0.068$, SE (robust) = 0.029, p-value = .021. Once Holm correction is performed for multiple comparisons (as in Figure 2), the p-value increases to 0.042.
14. See footnote 12 for the type of test used here.
15. Since one of the items going into our government performance scale is explicitly comparative ("Compared with other countries, the UK government has responded well to the coronavirus outbreak") there might be concerns that this item has an undue influence on our results. However, as can be seen from Figure 2 in the Supporting Information where we model the effect of experimental condition on each of the three individual items going into the scale separately, the direction and size of effect is in each case very similar. This suggests that none of the items are individually and disproportionally driving our results.
16. To rule out that our null finding was an artifact of low power owing to the number of factor levels, we re-ran this analysis with a binary partisanship variable

(Conservative vs. all other partisanship options), and did not see a significant effect (Figure 3 in the Supporting Information). This was not a pre-registered choice.

17. When no participants were dropped (following the reasoning for this deviation from the pre-registration plan provided in "Sampling procedures" above), the treatment effect for those at mean level of trust in the UK-only condition became significant: $B = 0.066$, SE (robust) = 0.029, p-value = .024.
18. See Table 5 in Supporting Information for full details, including standard errors and significance levels.

References

Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.

Ahlquist, J., Copelovitch, M., & Walter, S. (2020). The political consequences of external economic shocks: Evidence from Poland. *American Journal of Political Science*, *64*(4), 904–920. https://doi.org/10.1111/ajps.12503

Ahlstrom-Vij, K. (2021). Do we live in a "post-truth" era? *Political Studies*, *71*(2), 501–517. https://doi.org/10.1177/00323217211026427

Aldridge, R. W., Lewer, D., Katikireddi, S. V., Mathur, R., Pathak, N., Burns, R., Fragaszy, E. B., Johnson, A. M., Devakumar, D., Abubakar, I., & Hayward, A. (2020). Black, Asian and minority ethnic groups in England are at increased risk of death from COVID-19: Indirect standardisation of NHS mortality data. *Wellcome Open Research*, *5*(88), 88. https://doi.org/10.12688/wellcomeopenres.15922.2

Allen, W., & Ahlstrom-Vij, K. (2024). *Replication data for: Worst of the bunch: Visual comparative benchmarks change evaluations of government performance*. Harvard Dataverse. https://doi.org/10.7910/DVN/VMZIPF

Allen, W. L. (2023). The conventions and politics of migration data visualizations. *New Media & Society*, *25*(6), 1313–1334. https://doi.org/10.1177/14614448211019300

Allen, W. L., Bandola-Gill, J., & Grek, S. (2024). Next slide please: The politics of visualization during COVID-19 press briefings. *Journal of European Public Policy*, *31*(3), 729–755. https://doi.org/10.1080/13501763.2022.2160784

Amit-Danhi, E. R. (2022). Strategic Temporality: Information Types and Their Rhetorical Usage in Digital Election Visualizations. *International Journal of Communication*, *16*(June), 3354–3378. https://ijoc.org/index.php/ijoc/article/view/18500/3823

Arel-Bundock, V., Blais, A., & Dassonneville, R. (2021). Do voters benchmark economic performance? *British Journal of Political Science*, *51*(1), 437–449. https://doi.org/10.1017/S0007123418000236

Aytaç, S. E. (2018a). A response to arel-bundock, Blais, and Dassonneville. https://home.ku.edu.tr/~saytac/Aytac_response_ABD.pdf (14 November 2021).

Aytaç, S. E. (2018b). Relative economic performance and the incumbent vote: A reference point theory. *The Journal of Politics*, *80*(1), 16–29. https://doi.org/10.1086/693908

Aytaç, S. E. (2020). Do voters respond to relative economic performance? Evidence from survey experiments. *Public Opinion Quarterly*, *84*(2), 493–507. https://doi.org/10.1093/poq/nfaa023

Baele, S. J., Coan, T. G., & Sterck, O. C. (2017). Security through numbers? Experimentally assessing the impact of numerical arguments in security communication. *The British Journal of Politics & International Relations*, *20*(2), 459–476. https://doi.org/10.1177/1369148117734791

Bandola-Gill, J., Grek, S., & Tichenor, M. (2022). *Governing the sustainable development goals: Quantification in global public policy*. Palgrave Macmillan. https://library.oapen.org/bitstream/handle/20.500.12657/57299/978-3-031-03938-6.pdf?sequence=1&isAllowed=y

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., & Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

Becher, M., Brouard, S., & Stegmueller, D. (2023). Endogenous benchmarking and government accountability: Experimental evidence from the COVID-19 pandemic. *British Journal of Political Science*, *54*(2), 355–372. https://doi.org/10.1017/S0007123423000170

Bekkers, V., & Moody, R. (2014). *Visual culture and public policy: Towards a visual polity?* Routledge.

Besley, T. J., & Case, A. (1995). Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *The American Economic Review*, *85*(1), 25–45. https://www.jstor.org/stable/2117994

Bhandari, A., Larreguy, H., & Marshall, J. (2021). Able and mostly willing: An empirical anatomy of information's effect on voter-driven accountability in Senegal. *American Journal of Political Science*, *67*(4), 1040–1066. https://doi.org/10.1111/ajps.12591

Bisgaard, M. (2015). Bias will find a way: Economic-perceptions, attributions of blame, and partisan-motivated reasoning during crisis. *The Journal of Politics*, *77*(3), 849–860. https://doi.org/10.1086/681591

Bol, D., Giani, M., Blais, A., & Loewen, P. J. (2021). The effect of COVID-19 lockdowns on political support: Some good news for democracy? *European Journal of Political Research*, *60*(2), 497–505. https://doi.org/10.1111/1475-6765.12401

Broome, A., Homolar, A., & Kranke, M. (2018). Bad Science: International Organizations and the Indirect Power of Global Benchmarking. *European Journal of*

International Relations, *24*(3), 514–539. https://doi.org/10.1177/1354066117719320

Carlin, R. E., Hellwig, T., Love, G. J., Martínez-Gallardo, C., & Singer, M. M. (2021). When does the public get it right? The information environment and the accuracy of economic sentiment. *Comparative Political Studies*, *54*(9), 1499–1533. https://doi.org/10.1177/0010414021989758

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06, https://EconPapers.repec.org/RePEc:jss:jstsof:v:048:i06

Chanley, V. A., Rudolph, T. J., & Rahn, W. M. (2000). The origins and consequences of public trust in government: A time series analysis. *Public Opinion Quarterly*, *64*(3), 239–256. https://doi.org/10.1086/317987

Coppock, A. (2022). *Persuasion in parallel: How information changes minds about politics*. The University of Chicago Press.

Crepaz, M., & Arikan, G. (2021). Information disclosure and political trust during the COVID-19 crisis: Experimental evidence from Ireland. *Journal of Elections, Public Opinion, and Parties*, *31*(sup1), 96–108. https://doi.org/10.1080/17457289.2021.1924738

Damstra, A., Boukes, M., & Vliegenthart, R. (2020). To credit or to blame? The asymmetric impact of government responsibility in economic news. *International Journal of Public Opinion Research*, *33*(1), 1–17. https://doi.org/10.1093/ijpor/edz054

de Haan, Y., Kruikemeier, S., Lecheler, S., Smit, G., & van der Nat, R. (2018). When does an infographic say more than a thousand words? *Journalism Studies*, *19*(9), 1293–1312. https://doi.org/10.1080/1461670X.2016.1267592

Devine, D., Gaskell, J., Jennings, W., & Stoker, G. (2020). *Exploring trust, mistrust and distrust*. University of Southampton. https://static1.squarespace.com/static/5c21090f8f5130d0f2e4dc24/t/5e995ec3f866cd1282cf57b1/1587109577707/TrustGov+-+Trust+mistrust+distrust+-+20.04.2020.pdf

Devine, D., Gaskell, J., Jennings, W., & Stoker, G. (2021). Trust and the coronavirus pandemic: What are the consequences of and for trust? An early review of the literature. *Political Studies Review*, *19*(2), 274–285. https://doi.org/10.1177/1478929920948684

Devine, D., Valgarðsson, V., Smith, J., Jennings, W., Scotto di Vettimo, M., Bunting, H., & McKay, L. (2023). Political trust in the first year of the COVID-19 pandemic: A meta-analysis of 67 studies. *Journal of European Public Policy*, *31*(3), 657–679. https://doi.org/10.1080/13501763.2023.2169741

De Vries, C. E. (2018). *Euroscepticism and the future of European integration*. Oxford University Press.

De Vries, C. E., Bakker, B. N., Hobolt, S. B., & Arceneaux, K. (2021). Crisis signaling: How Italy's coronavirus lockdown affected incumbent support in other European countries. *Political Science Research and Methods*, *9*(3), 451–467. https://doi.org/10.1017/psrm.2021.6

De Vries, C. E., Hobolt, S. B., & Tilley, J. (2018). Facing up to the facts: What causes economic perceptions? *Electoral Studies*, *51*, 115–122. https://doi.org/10.1016/j.electstud.2017.09.006

Druckman, J. N. (2001). On the limits of framing effects: Who can frame? *The Journal of Politics*, *63*(4), 1041–1166. https://doi.org/10.1111/0022-3816.00100

Duch, R. M., & Stevenson, R. (2010). The global economy, competency, and the economic vote. *The Journal of Politics*, *72*(1), 105–123. https://doi.org/10.1017/S0022381609990508

Engebretsen, M., & Kennedy, H. (Eds.). (2020). *Data visualization in society*. Amsterdam University Press.

Falisse, J.-B., & McAteer, B. (2022). Visualising policy responses during health emergencies. Learning from the COVID-19 policy trackers. *Convergence: The International Journal of Research Into New Media Technologies*, *28*(1), 35–51. https://doi.org/10.1177/13548565211048972

Financial Times. (2020). Coronavirus tracked. https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938 (6 September 2020).

Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., Bellamy, D., Bibi, S., Bittaye, M., Clutterbuck, E. A., Dold, C., Faust, S. N., Finn, A., Flaxman, A. L., Hallis, B., Heath, P., Jenkin, D., Lazarus, R., & Makinson, R., Oxford COVID Vaccine Trial Group. (2020). Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: A preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet (London, England)*, *396*(10249), 467–478. https://doi.org/10.1016/S0140-6736(20)31604-4

Gadarian, S. K., Wallace Goodman, S., & Pepinsky, T. B. (2020). Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. https://doi.org/10.2139/ssrn.3562796 (9 September 2020).

Gilray, C. (2021). Performative control and rhetoric in Aotearoa New Zealand's response to COVID-19. *Frontiers in Political Science*, *3*, 86. https://doi.org/10.3389/fpos.2021.662245

Gomez, B. T., & Wilson, J. M. (2008). Political sophistication and attributions of blame in the wake of Hurricane Katrina. *Publius: The Journal of Federalism*, *38*(4), 633–650. https://doi.org/10.1093/publius/pjn016

Green, J., Edgerton, J., Naftel, D., Shoub, K., & Cranmer, S. J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, *6*(28), Article eabc2717. https://doi.org/10.1126/sciadv.abc2717

Hansard. (2020). *Prime minister's questions*. https://hansard.parliament.uk/commons/2020-05-13/debates/7F8336E1-EDA2-4AF8-9923-A65B8A9B3F7E/Engagements (4 September 2020).

Hansen, K. M., Olsen, A. L., & Bech, M. (2015). Cross-national yardstick comparisons: A choice experiment on a forgotten voter heuristic. *Political Behavior*, *37*(4), 767–789. https://doi.org/10.1007/s11109-014-9288-y

Hart, A., & Matthews, J. S. (2022). Unmasking accountability: Judging performance in an interdependent World. *The Journal of Politics*, *84*(3), 1607–1622. https://doi.org/10.1086/716971

Healy, A., & Malhotra, N. (2013). Retrospective voting reconsidered. *Annual Review of Political Science*, *16*(1), 285–306. https://doi.org/10.1146/annurev-polisci-032211-212920

Hellwig, T., & Coffey, E. (2011). Public opinion, party messages, and responsibility for the financial crisis in Britain. *Electoral Studies*, *30*(3), 417–426. https://doi.org/10.1016/j.electstud.2010.11.007

Hobolt, S. B. (2015). Public attitudes towards the Euro crisis. In O. Cramme, & S. B. Hobolt (Eds.), *Democratic politics in a European union under stress* (pp. 48–66). Oxford University Press.

Hobolt, S. B., & Tilley, J. (2014). *Blaming Europe: Attribution of responsibility in the European union*. Oxford University Press.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., & Scheibe, S. (2020). Multcomp: Simultaneous inference in general parametric models. https://CRAN.R-project.org/package=multcomp

Ipsos, MORI. (2020). *Coronavirus: Growing divisions over the UK government's response*. https://www.ipsos.com/sites/default/files/ct/news/documents/2020-05/coronavirus_-_growing_divisions_over_uk_government_response.pdf (10 July 2020).

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jennings, W. (2020). Covid-19 and the "rally-round-the flag" effect. https://ukandeu.ac.uk/covid-19-and-the-rally-round-the-flag-effect/ (10 September 2020).

Johansson, B., Hopmann, D. N., & Shehata, A. (2021). When the rally-around-the-flag effect disappears, or: When the COVID-19 pandemic becomes "normalized". *Journal of Elections, Public Opinion, and Parties*, *31*(sup1), 321–334. https://doi.org/10.1080/17457289.2021.1924742

Jørgensen, F., Bor, A., Lindholt, M. F., & Petersen, M. B. (2021). Public support for government responses against COVID-19: Assessing levels and predictors in eight Western democracies during 2020. *West European Politics*, *44*(5–6), 1129–1158. https://doi.org/10.1080/01402382.2021.1925821

Kayser, M., & Peress, M. (2012). Benchmarking across borders: Electoral accountability and the necessity of comparison. *American Political Science Review*, *106*(3), 661–684. https://doi.org/10.1017/S0003055412000275

Kayser, M., & Peress, M. (2021). Benchmarking across borders: An update and response. *British Journal of Political Science*, *51*(1), 450–453. https://doi.org/10.1017/S0007123418000625

Lazarus, J. V., Ratzan, S., Palayew, A., Billari, F. C., Binagwaho, A., Kimball, S., Larson, H. J., Melegaro, A., Rabin, K., White, T. M., & El-Mohandes, A. (2020).

COVID-SCORE: A global survey to assess public perceptions of government responses to COVID-19 (COVID-SCORE-10). *PLoS One*, *15*(10), Article e0240011. https://doi.org/10.1371/journal.pone.0240011

Lewis-Beck, M. S., & Stegmaier, M. (2000). Economic determinants of electoral outcomes. *Annual Review of Political Science*, *3*(1), 183–219. https://doi.org/10.1146/annurev.polisci.3.1.183

Lupia, A. (2015). *Why people know so little about politics and what we can do about it*. Oxford University Press.

Luskin, R. C. (1987). Measuring political sophistication. *American Journal of Political Science*, *31*(4), 856–899. https://doi.org/10.2307/2111227

Lyons, J., & Jaeger, W. P. (2014). Who do voters blame for policy failure? Information and the partisan assignment of blame. *State Politics and Policy Quarterly*, *14*(3), 321–341. https://doi.org/10.1177/1532440014537504

Malhotra, N., & Kuo, A. G. (2008). Attributing blame: The public's response to Hurricane Katrina. *The Journal of Politics*, *70*(1), 120–135. https://doi.org/10.1017/S0022381607080097

Martinez-Bravo, M., & Sanz, C. (2023). *Trust and accountability in times of pandemics*. Documentos de Trabajo. https://doi.org/10.53479/29471

Martini, M., & Walter, S. (2023). Learning from precedent: How the British brexit experience shapes nationalist rhetoric outside the UK. *Journal of European Public Policy*, *31*(5), 1231–1258. https://doi.org/10.1080/13501763.2023.2176530

McNaughton, C. D., Cavanaugh, K. L., Kripalani, S., Rothman, R. L., & Wallston, K. A. (2015). Validation of a short, 3-item version of the subjective numeracy scale. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *35*(8), 932–936. https://doi.org/10.1177/0272989X15581800

Mishler, W., & Rose, R. (2001). What are the origins of political trust? Testing institutional and cultural theories in post-communist societies. *Comparative Political Studies*, *34*(1), 30–62. https://doi.org/10.1177/0010414001034001002

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How Conditioning on Post-treatment Variables Can Ruin Your Experiment and What to Do About It. *American Journal of Political Science*, *62*(3), 760–775. https://doi.org/10.1111/ajps.12357

Mueller, J. E. (1970). Presidential popularity from Truman to Johnson. *American Political Science Review*, *64*(1), 18–34. https://doi.org/10.2307/1955610

Nadeau, R., Lewis-Beck, M. S., & Bélanger, É. (2012). Economics and elections revisited. *Comparative Political Studies*, *46*(5), 551–573. https://doi.org/10.1177/0010414012463877

Narlikar, A., & Sottilotta, C. E. (2021). Pandemic narratives and policy responses: West European governments and COVID-19. *Journal of European Public Policy*, *28*(8), 1238–1257. https://doi.org/10.1080/13501763.2021.1942152

Our World in Data. (2024). Coronavirus (COVID-19) deaths. https://ourworldindata.org/covid-deaths (1 May 2023).

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*(May), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pentzold, C., Fechner, D. J., & Conrad, Z. (2021). "Flatten the curve": Data-driven projections and the journalistic brokering of knowledge during the COVID-19 crisis. *Digital Journalism*, *9*(3), 1–24. https://doi.org/10.1080/21670811.2021.1950018

Petersen, M. B., Bor, A., Jørgensen, F., & Lindholt, M. F. (2021). Transparent communication about negative features of COVID-19 vaccines decreases acceptance but increases trust. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(29), Article e2024597118. https://doi.org/10.1073/pnas.2024597118

Powell, G. B., & Whitten, G. D. (1993). A cross-national analysis of economic voting: Taking account of the political context. *American Journal of Political Science*, *37*(2), 391–414. https://doi.org/10.2307/2111378

R Core Team. (2018). *R: A language and environment for statistical computing*. R Core Team.

Roberts, S. (2020). *Flattening the coronavirus curve*. The New York Times. https://www.nytimes.com/article/flatten-curve-coronavirus.html (11 October 2020).

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2021). Coronavirus pandemic (COVID-19). *Our World in Data*. https://ourworldindata.org/coronavirus (31 October 2021).

Sevi, S., Aviña, M. M., Péloquin-Skulski, G., Heisbourg, E., Vegas, P., Coulombe, M., Arel-Bundock, V., Loewen, P. J., & Blais, A. (2020). Logarithmic versus linear visualizations of COVID-19 cases do not affect citizens' support for confinement. *Canadian Journal of Political Science*, *53*(2), 385–390. https://doi.org/10.1017/S000842392000030X

Shin, J., & Park, B. B. (2022). Benchmarking the pandemic: How do citizens react to domestic COVID-19 conditions compared to other countries? *Journal of Elections, Public Opinion, and Parties*, *34*(2), 1–21. https://doi.org/10.1080/17457289.2022.2120887

Spälti, A. K., Lyons, B., Mérola, V., Reifler, J., Stedtnitz, C., Stoeckel, F., & Szewach, P. (2021). Partisanship and public opinion of COVID-19: Does emphasizing Trump and his administration's response to the pandemic affect public opinion about the coronavirus? *Journal of Elections, Public Opinion, and Parties*, *31*(sup1), 145–154. https://doi.org/10.1080/17457289.2021.1924749

Tilley, J., & Hobolt, S. B. (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. https://doi.org/10.1017/S0022381611000168

UK Government. (2023). Coronavirus (COVID-19) in the UK. https://coronavirus.data.gov.uk/details/deaths (26 February 2024).

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

World Health Organization. Coronavirus disease (COVID-19) advice for the public: Mythbusters. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (20 September 2020).

Worthy, B. (2010). More open but not more trusted? The effect of the freedom of information act 2000 on the United Kingdom central government. *Governance*, *23*(4), 561–582. https://doi.org/10.1111/j.1468-0491.2010.01498.x

Yagci, A. H., & Oyvat, C. (2020). Partisanship, media and the objective economy: Sources of individual-level economic assessments. *Electoral Studies*, *66*(August), 102135. https://doi.org/10.1016/j.electstud.2020.102135

Young, M. L., Hermida, A., & Fulda, J. (2018). What makes for great data journalism? *Journalism Practice*, *12*(1), 115–135. https://doi.org/10.1080/17512786.2016.1270171

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1–17. https://doi.org/10.18637/jss.v011.i10

Author Biographies

William L. Allen is a British Academy Postdoctoral Research Fellow in the Department of Politics and International Relations at the University of Oxford, and a Non-Stipendiary Research Fellow of Nuffield College. His research focuses on how and why people engage with information about economic and political issues, and what this means for political behavior and policymaking.

Kristoffer Ahlstrom-Vij is a quantitative social scientist, and formerly Head of the Department of Philosophy at Birkbeck University of London as well as a Research Fellow at the National Institute of Economic and Social Research (NIESR). His research interests involve the relationship between knowledge and political preferences.