AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

The Variable Persuasiveness of Political Rhetoric

Jack Blumenau and Benjamin E. Lauderdale University College London

Abstract: *Which types of political rhetoric are most persuasive? Politicians make arguments that share common rhetorical elements, including metaphor, ad hominem attacks, appeals to expertise, moral appeals, and many others. However, political arguments are also highly multidimensional, making it difficult to assess the relative persuasive power of these elements. We report on a novel experimental design which assesses the relative persuasiveness of a large number of arguments that deploy a set of rhetorical elements to argue for and against proposals across a range of UK political issues. We find modest differences in the average effectiveness of rhetorical elements shared by many arguments, but also large variation in the persuasiveness of arguments of the same rhetorical type across issues. In addition to revealing that some argument-types are more effective than others in shaping public opinion, these results have important implications for the interpretation of survey-experimental studies in the field of political communication.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/POMIFD.

Politicians invest time and effort in crafting arguments to present to voters, and the arguments that they make often deploy common rhetorical elements. Regardless of the specific policy at stake, politicians can draw on endorsements from relevant authorities, emphasize a moral rationale, carefully articulate costs and benefits, impugn the motives of opposition actors, present evidence from historical or other countries' experiences, and so on. While interest in rhetorical strategies has been sustained over the course of millennia (Aristotle 2004; Charteris-Black 2011; *Rhetorica ad Herennium* 2022; Riker 1990), and more recent work has begun to test the efficacy of different communication strategies (Bos, Van der Brug, and de Vreese 2013; Boudreau and MacKenzie 2014; Bougher 2012; Hameleers, Bos, and de Vreese 2017; Hameleers and Schmuck 2017; Jerit 2009; Jung 2020; Lau, Sigelman, and Rovner 2007; Loewen, Rubenson, and Spirling 2012; Nelson 2004; Schlesinger and Lau 2000; Thibodeau and Boroditsky 2011), making general statements about the *relative* performance of particular rhetorical strategies is difficult because arguments are so highly multidimensional. Arguments deploy common elements, but they also vary in many other ways that might make certain strategies more effective in some implementations than others. As a result, empirical research has rarely moved beyond demonstrating nonzero effectiveness of specific types of argument that politicians employ in particular domains. As a consequence, "scholars still understand little about the factors that shape argument strength" (Arceneaux 2012, 272).

Why is it important to determine whether some types of argument are more successful than others? Classical critiques suggest that political rhetoric is generally and inherently damaging to democracy because it prioritizes emotion and passion over reason and inhibits rational deliberation between citizens (Elster 1998). However, recent work in normative political theory that attempts to "rehabilitate rhetoric" (Chambers 2009; Dryzek 2010) suggests that while rhetoric may not be damaging per se, specific *forms* of rhetoric—particularly when used to communicate "vapid and vacuous" statements rather

Jack Blumenau is Associate Professor of Political Science and Quantitative Research Methods, Department of Political Science, University College London, 36-38 Gordon Square, London WC1H 9QU, United Kingdom (j.blumenau@ucl.ac.uk). Benjamin E. Lauderdale is Professor of Political Science, Department of Political Science, University College London, 36-38 Gordon Square, London WC1H 9QU, United Kingdom (b.lauderdale@ucl.ac.uk).

We thank Adam McDonnell at YouGov for implementing the survey designs. We also thank the *AJPS* editors and anonymous reviewers, as well as colleagues at University College London, the London School of Economics, Princeton University, and the Annual Conference of the Political Methodology Group of the Political Studies Association for their helpful feedback and comments. This work was supported in part by an ESRC Future Research Leaders grant (ES/N016297/2, Blumenau).

American Journal of Political Science, Vol. 00, No. 0, xxxx 2022, Pp. 1–16

DOI: 10.1111/ajps.12703

than substantive policy information—should still be viewed as a threat to deliberative ideals (Chambers 2009, 337). If voters consistently respond to arguments that are low in informational content but rich in bombast and élan, we might worry that the quality of deliberation has fallen. By contrast, if voters are more consistently persuaded by arguments that reference relevant factual information and expert authority, we might have less concern. Moreover, to the degree that voters' susceptibility to persuasive rhetoric undermines the idea that individuals hold stable and coherent political preferences (Bartels 2003), we believe it is important to establish *which forms* of rhetoric are most likely to shift public opinion across issues. To fully appreciate the implications of rhetorical persuasion for democratic theory, we require a stronger understanding of "when and why some of those strategies work better than others" (Bartels 2003, 68).

In this article, we provide the first quantitative evaluation of the relative effectiveness of a large number of different rhetorical elements across a large number of political issues by introducing a new experimental design and associated modeling approach. We examine types of arguments frequently made in contemporary British politics, and especially in speeches delivered in the UK parliament. The rhetorical elements that we identify relate to ongoing debates in diverse literatures in political communication, and are relevant to domestic politics in many countries. Our main experiment tests 336 individual arguments that use one of 14 distinct rhetorical elements to make arguments on each side of 12 policy issues in the United Kingdom. We present pairs of these arguments to survey respondents and ask them to assess which of the pair is most persuasive. We then use the distribution of responses to generate estimates of the relative persuasiveness of each of the arguments and, in turn, of the average persuasiveness of each of the rhetorical elements. A central virtue of our design is that, by presenting many implementations of each element, we are able to draw inferences about the relative effectiveness of different rhetorical strategies averaged across different political issues.

In addition to being a study of argument types and their relative persuasiveness, this article is also a methodological argument for a different sort of experimental design. Recent meta-analyses of persuasion field experiments (Kalla and Broockman 2018) and online advertising experiments (Coppock, Hill, and Vavreck 2019) move beyond merely collecting existing study results toward fielding multiple similar experiments for the purpose of pooling evidence from them. Our design takes this logic much further. Researchers using survey experiments seldom want to test the effects of particular treatment texts on particular survey prompts. Rather, they typically want to make broader claims about a *latent treatment* (Grimmer and Fong 2022) or treatment type, of which a treatment text is just one implementation. Many of the latent treatments that researchers wish to assess are likely to have variable effects across specific implementations. If we are interested in the *type* of treatment, rather than the specific treatment text, using many implementations rather than few or one should not wait for a meta-analysis of a mature research literature.

A traditional objection to this is that we would need to collect far larger samples to test many implementations of a latent treatment type. However, once we recognize that we are far less interested in the effects of specific treatment implementations than the *distribution* of such effects across implementations, we can use multilevel modeling to estimate this distribution using a large number of implementations, each of which would be statistically underpowered if analyzed alone. In addition to reducing the risk that our conclusions about the latent treatment types will be confounded by the idiosyncrasies of single implementations, we illustrate how this approach enables postexperimental checks related to specific confounding concerns.

Our main substantive results reinforce the value of these methodological innovations. We find that there are modest average differences between different rhetorical element types. One of the strongest rhetorical elements in our experiment is *appeals to authority*—that is, arguments that seek support for an issue by reporting the view of an entity with relevant subject area expertise. The role of expertise in political debate became a prominent issue in UK politics during the Brexit referendum in 2016 when a leading figure in the Leave campaign declared that the public "have had enough of experts."[1] Our results suggest that, despite this view, making appeals to authority remains among the most persuasive ways to argue about political issues. By contrast, the weakest arguments, on average, are those that employ ad hominem attacks and those that rely on *metaphor and imagery* to win support for a policy stance. Although empirical evidence on the efficacy of negative attacks in political communication is mixed (Lau, Sigelman, and Rovner 2007), recent studies argue that the use of metaphor can be central to successful political campaigning (Charteris-Black 2011) and a major determinant of the ways that individuals reason about politics (Thibodeau and Boroditsky 2011). Our results build on both of these literatures and suggest that when compared to many other common forms of political rhetoric, arguments of these types are relatively unpersuasive in the eyes of the UK public, *at least on average*.

[1] See Mance (2016).

However, and in some sense more importantly, we find that the heterogeneity in the effectiveness of specific implementations of these rhetorical elements is much larger than these average differences. While *appeals to authority* are more persuasive than other rhetorical styles on average, some appeals of this sort are still among the weakest arguments we test. Similarly, arguments that rely on making *comparisons to other countries* feature in the lists of the most and least persuasive in our experiment, depending on the specific implementation and issue. This finding represents an important lesson for the interpretation of existing studies of rhetorical effectiveness in political communication, a large number of which are based on experiments that relate to single policy issues. While it is not novel to observe that the effects of particular experimental implementations may not generalize to other domains, we directly quantify the substantial variance of the effects of the same treatment types across issues.

Ultimately, our goal is to understand which types of argument induce voters to support or oppose policy proposals on different issues. However, *persuasion* of this sort is different from the self-reported judgments of argument *persuasiveness* that we elicit in our experiment (Graham and Coppock 2021; Vavreck 2007). To address this concern, we conduct a separate, out-of-sample validation experiment. We find that respondents' evaluations of which arguments are more *persuasive* in our initial experiment strongly predict the direction and magnitude of those arguments' ability to *persuade* different respondents to actually change their stated attitudes in the validation. The validation demonstrates large persuasion effects on average, but we again observe large variation in these treatment effects across policy issues. Therefore, in addition to providing an important check on the validity of our main experimental design and measurement strategy, the validation also reinforces our central methodological argument. Argument quality varies substantially, and researchers should exercise caution when generalizing the results of studies in specific policy areas to different issue domains.

Rhetoric, Persuasion, and Public Opinion

Canonical work in the literature takes a broad view of what constitutes political rhetoric, seeing it as a "range of methods for persuading others" (Charteris-Black 2011, 13). Politicians' arguments often share common rhetorical elements that are thought to be one source of their persuasive appeal, and we share the understanding of Atkins and Finlayson (2013, 161) that analyses of political rhetoric should focus "on the varied kinds of proof or justification found in political argument." Several existing typologies partition political arguments into a number of distinct rhetorical categories (e.g., Aristotle 2004; Charteris-Black 2011; Finlayson 2007), but—as we describe below—our focus is on argument types that arise regularly in UK politics.

Research into political rhetoric is not always described as such, but one goal of a large body of public opinion research is to measure the persuasive effects of different forms of political argument. The core conceptual focus of research in this field is whether (and to what degree) a given rhetorical element can persuade citizens to change their political views. For instance, though politicians may construct very different metaphors to argue about the economy (Barnes and Hicks 2019), crime (Thibodeau and Boroditsky 2011), and healthcare (Schlesinger and Lau 2000), it might be the use of metaphor itself that is "essential to their persuasiveness" (Charteris-Black 2011, 2). Existing work has considered the effects of a wide range of rhetorical elements on public opinion, including populist rhetoric (Atkins and Finlayson 2013; Bos, Van der Brug, and de Vreese 2013; Hameleers, Bos, and de Vreese 2017; Hameleers and Schmuck 2017); negative or ad hominem attacks (Lau, Sigelman, and Rovner 2007); morality- and values-based appeals (Jung 2020; Nelson 2004); appeals based on expected costs and benefits of policy (Jerit 2009; Riker 1990); and the use of expert cues and endorsements (Atkins and Finlayson 2013; Boudreau and MacKenzie 2014; Dewan, Humphreys, and Rubenson 2014).

Similarly, the literature on framing effects asks whether strategic language use by elites can change the factors that are relevant to voters' evaluations of policy options. Existing research in this area considers frames that emphasize free speech concerns (Nelson, Clawson, and Oxley 1997), fiscal cost-benefit considerations (Leeper and Slothuus 2018, 15), and the importance of civil liberties (Chong and Druckman 2010), among others. Though these studies are often concerned with how political communications are portrayed in the mass media, they all engage with the idea that politicians can persuade voters to endorse particular policy options by using language strategically.

Our study addresses three limitations of the existing literature on rhetoric and persuasion. First, the cumulative evidence from these studies suggests that elite communication can substantially shift public opinion, and several authors express concern that such results imply that citizens do not hold stable and well-formed

preferences (Disch 2011; Druckman 2004). Others have argued that there is an important role for rhetoric in the process of democratic deliberation (Dryzek 2010), and those certain forms of rhetoric are more defensible than others (Chambers 2009). Chambers (2009, 328), for example, emphasizes that it is not rhetoric per se that is problematic, but specifically *plebiscitary* rhetoric—populist appeals divorced from factual merits—that represents a "threat to deliberation." By contrast, less problematic is *deliberative* rhetoric, which "makes people think, it makes people see things in new ways, it conveys information and knowledge, and it makes people more reflective" (Chambers 2009, 335). A key goal for empirical studies, then, should be to determine whether different forms of rhetoric are differentially persuasive.

Unfortunately, the existing evidence on rhetoric and persuasion, which comes predominantly from survey experiments, provides little information regarding such comparisons. In almost all the papers cited above (and many others not cited) the persuasiveness of the relevant style or frame of interest—populism, metaphor, morality, and so on—is evaluated in the context of vignette experiments where a treatment text containing the relevant element is contrasted with a control condition that does not include that element.[2] We are not the first to observe that *comparisons* of persuasiveness between the element of interest and other plausibly applicable rhetorical elements are very rare (Chong and Druckman 2007b, 638; Sniderman and Theriault 2004, 141). Thus, our first contribution is to provide novel evidence about which of a relatively large number of types of political rhetoric are more or less effective for shaping public opinion.

Second, the overwhelming majority of survey experiments that estimate the causal effects of persuasive speech do so in the context of a single issue. Existing work on external validity in survey experiments has explored whether effects estimated from convenience samples match those from representative samples (Berinsky, Huber, and Lenz 2012; Coppock 2019) and whether experimental findings are replicated in comparable real-world settings (Barabas and Jerit 2010; Bechtel et al. 2015). However, these external validity concerns are distinct from the idea that effects detected in an experiment on one issue may not generalize to a broader population of political issues for which politicians might use these types of rhetoric. As Druckman (2004, 685) suggests, "scholars need to carefully consider the context under study—perhaps, to an even greater extent than the population."

[2]See, for example, Arceneaux (2012), Bos, Van der Brug, and de Vreese (2013), Jerit (2009), Jung (2020), and Nelson (2004).

Some existing research (Hopkins and Mummolo 2017; Lecheler, de Vreese, and Slothuus 2009) suggests that the persuasive effects of different frames vary across policy issues, and it seems plausible a priori that certain types of rhetoric may be more appropriate for certain policy issues. For instance, are the legitimizing effects of populist rhetoric the same for issues relating to nuclear power (Bos, Van der Brug, and de Vreese 2013) as they are for immigration? Are loss aversion arguments equally persuasive on economic issues as they appear to be on public health issues (Arceneaux 2012)? Are rhetorical statements that make reference to cost-benefit considerations as influential when applied to issues of education as they are to issues of welfare (Jerit 2009)? Of course, many studies in this literature make nuanced arguments about rhetorical effectiveness, paying close attention to the conditions under which different strategies are likely to be persuasive. Nevertheless, understanding whether some rhetorical elements are predictably more or less effective when considered across multiple policy issues remains an important and open question. Our second contribution, therefore, is to provide evidence of the distribution of effectiveness of rhetorical elements across a wide range of political issues.

Third, our approach also helps us to overcome a methodological problem that is common to vignette-style experiments that use single-text treatments. Grimmer and Fong (2022) argue that latent treatments that are of interest to the researcher often co-occur with other textual features in experimental treatment texts. When this is true, effects estimated from such texts cannot necessarily be attributed to the latent treatment, as they might reflect instead the effects of these other correlated features.

In our design, which is similar to the design in Grimmer and Fong (2022), we provide *several* texts per latent treatment, thereby allowing background features that might confound our latent concepts of interest to vary. If background features vary independently of the concept of interest, then researchers can average over the effects of these separate treatments and attribute the average effect to the latent concept. Even if these potentially confounding background features do not vary independently of the concept of interest, having multiple treatment texts means that we are able to statistically control for any measurable confounding features of those texts. In combination, these aspects of our design mean that we can be much more confident that the treatment effects that we estimate in our experiment are attributable to the latent concepts (i.e., rhetorical elements) that motivate our study.

We build on the intuition and formalization presented in Grimmer and Fong (2022), but our study differs from theirs in terms of research design and empirical application. First, the central estimand for Grimmer and Fong is the average marginal component effect of their latent treatments, which they estimate using linear regression. By contrast, we illustrate how to use a multilevel modeling approach to characterize the *average* effect of each of our latent treatments as well as the *distribution* of the treatment effects across text implementations. Quantifying the variation in treatment effects across implementations provides important information about the generalizability of findings from existing single-implementation studies. Second, whereas Grimmer and Fong provide evidence of the effectiveness of using multiple text-based implementations in the context of a single example, we apply this idea to a much larger set of latent treatments in the field of political persuasion and document substantively important findings about the heterogeneity in argument strengths of different types.

Experimental Design

We start by distinguishing between three concepts that are central to the structure of our experimental design: policy issues, rhetorical elements, and arguments.

A *policy issue* refers to an issue that is subject to some level of political debate, where the government could plausibly take action. In our setting, we focus on 12 policy issues in contemporary British politics: "Building a third runway at Heathrow," "Closing large retail stores on Boxing Day," "Extending the Right to Buy," "Extension of surveillance powers in the UK," "Fracking in the UK," "Nationalization of the railways in the UK," "Quotas for women on corporate boards," "Reducing the legal restrictions on cannabis use," "Reducing university tuition fees," "Renewing Trident," "Spending 0.7% of GDP on overseas aid," and "Sugar tax in the UK." In deciding which policies to include, we focused on identifying those where there were clear political disagreements, both among politicians and the public, but where these divisions were not among the highest-profile issues in British politics.

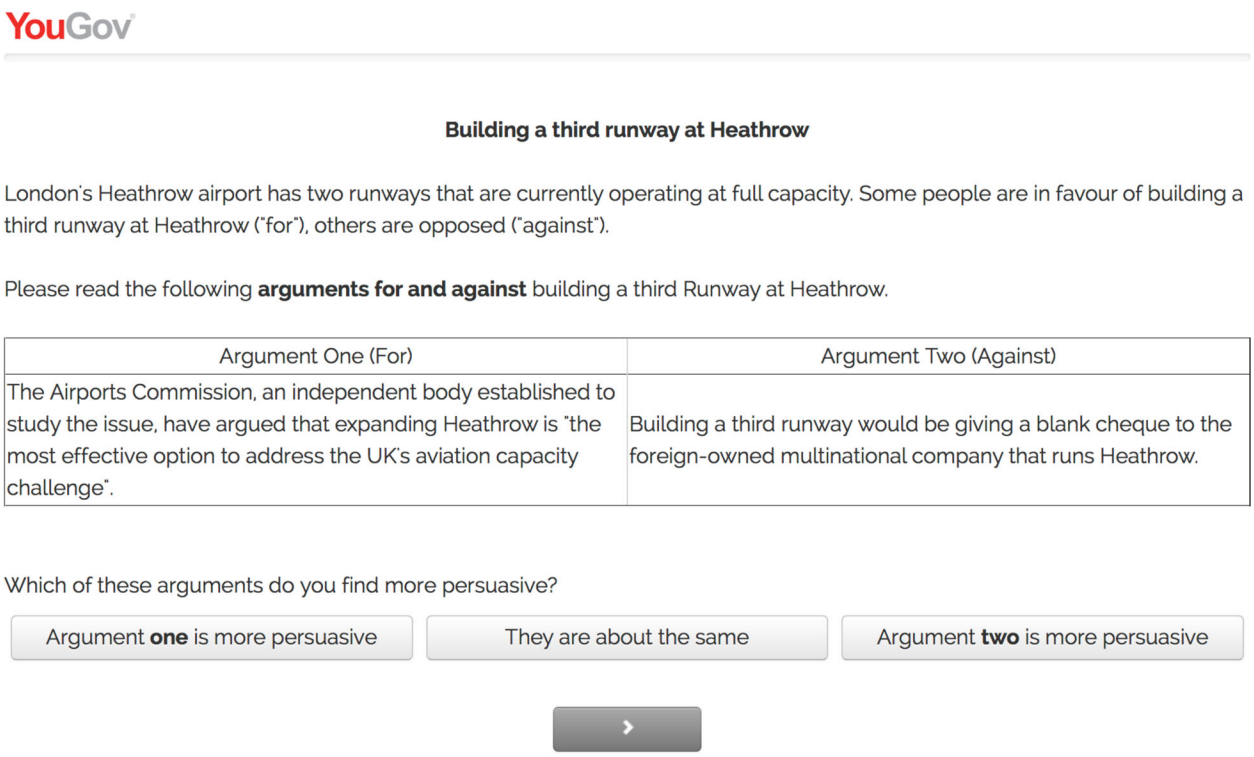
A *rhetorical element* is a feature of political argument that is used to emphasize the desirability or undesirability of a given policy. We based our categorization of rhetorical elements on close reading of contemporary political debates. We began with a list of possible rhetorical categories and then expanded and refined our categorization by reading through transcripts of debates in the UK House of Commons and House of Lords that related to the issues defined above. Sourcing our arguments from parliamentary debates is helpful for situating our study in the context of real-world politics, and it is consistent with calls to study "political arguments as they take place 'in the wild'" (Finlayson 2007, 552). These debates provide a large repository of arguments about specific policy areas, which tend to mirror those used by UK politicians in public speeches outside of parliament. The set of rhetorical elements that we evaluate, which was not intended to be exhaustive, is given in Tables 1–12 in the appendix.[3] While our primary goal is to quantify the persuasiveness of rhetorical appeals used in contemporary politics, our design is amenable to any arbitrary categorization of arguments into types so long as the researcher is able to write multiple implementations of the same treatment concepts.

An *argument* is a text that makes a case in favor of or against a specific policy. While real-world arguments sometimes include multiple rhetorical elements, for the purposes of our experiment we designed arguments that used a single element from the typology that we developed. A consequence of this decision is that our experiment is unable to evaluate whether certain rhetorical elements are more or less effective when used in conjunction with other elements. Although interaction effects between elements are possible—and in some cases likely—we focus here on establishing the relative persuasiveness of our rhetorical elements when considered individually. For each policy issue, we wrote two separate arguments for each of the rhetorical elements: one arguing in favor of the policy, and one arguing against. This results in $14 \times 12 \times 2 = 336$ separate arguments that are the basic treatments in our experiment.

To ensure the arguments we used resembled the types of argument used by politicians in the United Kingdom, we searched through the transcripts of UK parliamentary debates that pertained to the policy issues outlined above. From these debates, we extracted sentences and paragraphs that corresponded to our rhetorical elements, and then we edited these texts into the form we use in the experiment. In the appendix (pp. 20–34) we present all 336 arguments, and for many of the sentences we provide hyperlinks to the relevant source documents. When it was not possible to identify an example of our rhetorical styles in the texts of the Commons' debates on a particular issue, we wrote argu-

[3]We considered further element types that feature in UK debate—such as "examples of personal narrative" and "appeals to freedom"—but found it too difficult to write treatments for them across all issues in the experiment.

FIGURE 1 Survey Prompt for Experiment 1



Note: Prompt fielded to 3,317 respondents in June 2019.

ments of our own, making the texts as similar as possible in style to those based on politicians' speeches.[4]

Survey Instrument

We use these arguments as the basis of a forced-choice experiment that was fielded by YouGov to their UK online panel in June 2019. Following an introduction screen describing the task, respondents were presented with two arguments pertaining to a particular policy issue and asked which of the two arguments they found more persuasive.[5] Policy issues were sampled from the full set of 12 policy issues. For the selected policy, we then randomly sampled whether a respondent was presented with two arguments "in favor" of that policy (25% of responses), two arguments "against" that policy (25% of responses), or one argument "in favor" and one argument "against" (50% of responses).[6] We collected responses on four randomly selected issues from each of 3,317 respondents, giving us a total of 13,268 observations. An example prompt is given in Figure 1.

As the wording of the survey prompt clearly reflects, this experiment assesses "persuasiveness" rather than "persuasion." That is, we look at *self-reported* assessments of arguments by respondents rather than the treatment effects of different arguments on respondents' own positions. Survey respondents tend to overestimate the effects of political stimuli on their own behavior and attitudes (Graham and Coppock 2021; Vavreck 2007), so we might be concerned that this approach will lead to overestimates of the variation in argument strengths. We address this issue by implementing an out-of-sample validation study (described below) where we check whether the arguments that respondents say are more persuasive are, in fact, better able to persuade people to endorse different policy positions.

[4]In our analysis below, we show that we were no better or worse than UK politicians at writing persuasive political arguments.

[5]We use a paired-choice design for two reasons. First, by presenting respondents with competing arguments, our experiment more closely approximates the ways in which voters are exposed to political debate in the real world—a view that is shared by others (Arceneaux 2012, 271; Chong and Druckman 2007a, 102). Second, paired-profile designs of this sort are more successful, relative to single-profile designs, at generating estimates that replicate known real-world behavioral benchmarks (Hainmueller, Hangartner, and Yamamoto 2015).

[6]In the appendix (pp. 14–15), we show that "same side" and "different side" comparisons result in argument rankings that correlate at 0.81, which indicates that we can get nearly the same information from the different types of comparisons.

Few prior studies provide comparisons of the effectiveness of different rhetorical elements, or comparisons of the effectiveness of the same element across multiple issues, so we did not set out to test specific expectations about the relative persuasiveness of the elements we include in the experiment. Rather, we see our experiment as a measurement exercise that allows us to try to decompose the elements of a persuasive appeal that make it more or less effective. Our design, combined with the modeling framework outlined in the next section, allows us to estimate two key quantities relevant to this goal: the average (self-reported) persuasiveness of a wide variety of rhetorical elements, and the variation in persuasiveness of those elements across issues.

The Relative Persuasiveness of Rhetorical Elements

Modeling Persuasiveness

Our design generates responses that specify "winners" from a pairwise competition between arguments, with the possibility of ties.[7] Overall, we have J arguments, which we denote with $j = 1, \ldots, J$, and which we present to respondents, indexed as $i = 1, \ldots, N$, in paired comparisons. Our modeling task is to infer the efficacy of particular types of argument given the results of the pairwise contests.

Our experiment results in an ordered response variable with three categories:

$$Y_i \in \begin{cases} 1 = \text{Argument 2 is more persuasive;} \\ 2 = \text{About the same;} \\ 3 = \text{Argument 1 is more persuasive.} \end{cases} \quad (1)$$

To model this outcome, we adopt a variation on the Bradley-Terry model for paired comparisons (Bradley and Terry 1952) where we model the log-odds that argument j beats argument j' in a pairwise comparison as follows:

$$log\left[\frac{P(Y_{j,j'} \leq k)}{P(Y_{j,j'} > k)}\right] = \theta_k + \alpha_j - \alpha'_j \quad (2)$$

where θ_k is the cutpoint for response category k and each argument j is described by a single "strength" parameter α_j. The strength parameter for a given argument, α_j, increases in the number of comparisons j "wins" against other arguments, and also in the strength of the arguments that j defeats. The intuition behind these parameters is straightforward: the stronger argument j is relative to argument j', the higher the probability that argument j beats argument j' in a pairwise comparison.

If we had only a few arguments to test and a large number of responses involving each one, then we could simply use this as our full model specification and interpret the α_j parameters.[8] However, our primary quantity of interest is not the strength of individual arguments, but rather the *distribution* of the strength of arguments α_j for each of the 14 rhetorical elements. That is, we are interested in modeling how the strength of arguments varies as a function of rhetorical features that appear in those arguments. We therefore specify a hierarchical model for the α_j parameters.

Where $e(j) \in 1, \ldots, 14$ is the rhetorical element present in argument j, and $p(j) \in 1, \ldots, 12$ is the policy issue that the argument is about, and $s(j) \in 1, 2$ is the side of the issue that j argues for, we model the argument effects at a second level using a model of the following form:

$$\begin{aligned} \alpha_j &= \delta_{p(j),s(j)} + \mu_{e(j)} + \nu_j \\ \mu_e &\sim N(0, \omega) \\ \nu_j &\sim N\left(0, \sigma_{e(j)}\right). \end{aligned} \quad (3)$$

We assume a baseline effectiveness of arguments on the "for" versus the "against" side of each issue via the $\delta_{p(j),s(j)}$ parameters. These parameters separate the relative self-reported persuasive power of arguments from the degree to which respondents tend to agree with the side of the issue on which that argument appears. Note that given the way the α parameters enter Equation (2), these parameters cancel in the case where both arguments in the pairwise comparison are on the same side of an issue. The next set of parameters $\mu_{e(j)}$ captures the average effect of each of our rhetorical elements. The final set of parameters are the ν_j, which are argument-specific "residuals" that characterize the distribution of argument-level effects around the element-type average. We estimate separate variance parameters for each element ($\sigma_{e(j)}$).[9] We estimate the model using Stan (Carpenter et al. 2017).[10]

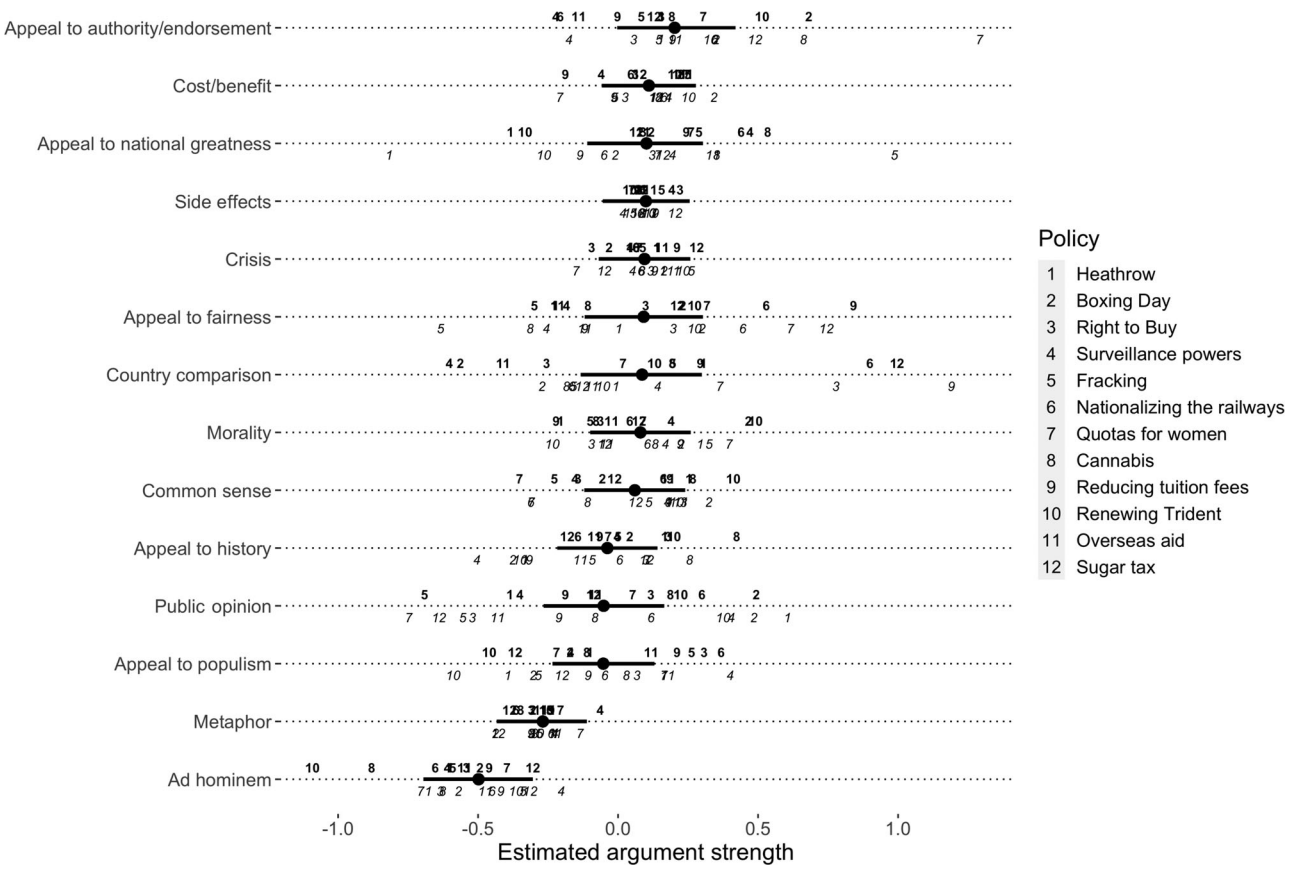
[7] An alternative would be to ask respondents to *rate* different arguments on a common scale. Such an approach might result in more fine-grained information, but it would come at the cost of decreased interpersonal comparability. The core of our approach would nonetheless apply to data from a rating task, with an appropriate multilevel model for an interval-level outcome.

[8] Each of the 336 individual treatments appears in an average of 79 pairwise comparisons in our data (s.d. = 8.6).

[9] To identify the relative scale, $\delta_{p(j),1} = 0$ for all "against" arguments, and $\delta_{p(j),2}$ is estimated with a uniform prior for all "for" arguments. We use uniform priors on the ω and σ_e parameters as well.

[10] We used four chains of 1,000 iterations, after a 200-iteration burn-in. We present convergence diagnostics in the appendix (pp. 17–19).

FIGURE 2 Treatment Strengths for Each Rhetorical Element and Each Individual Argument



Note: The figure shows the treatment strengths for each rhetorical element (points and intervals) and each individual argument (numbers) as estimated from the first experiment and associated multilevel model. Arguments on the "for" side of each issue are in bold and above the relevant element line. Arguments on the "against" side of each issue are in italics and below the relevant element line.

The hierarchical model described by Equation (3) distinguishes our approach from work on Canadian referendum arguments by Loewen, Rubenson, and Spirling (2012), who also ask survey respondents to compare pairs of political arguments and estimate the persuasive power of those arguments via a structured Bradley-Terry model. In essence, Loewen, Rubenson, and Spirling (2012) implement a version of the Equation (2), whereas we use Equation (2) as the basis of a hierarchical model that better captures our primary quantities of interest. While the two modeling approaches are similar at the argument level, our hierarchical model allows us to describe the distribution of element strengths across issues.

As we illustrate in the appendix (pp. 3–8), it is possible to recover very similar estimates for individual argument strengths—and nearly identical estimates for average rhetorical element strengths—using standard regression methods. We adopt a multilevel framework because the variance across treatments is a core quantity of interest, but we acknowledge that other analytic options are available for this kind of experiment.

Results

We present the main results from our model in Figure 2.[11] The figure shows the estimated average strength for each of our 14 rhetorical elements (μ_e) as well as for each of the 336 individual arguments ($\mu_{e(j)} + \nu_j$) that we include in the experiment. Bold numbers, above each dotted line, indicate arguments on the "for" side of the relevant issue, and italic numbers, below each line, indicate arguments on the "against" side of the issue. The numbers themselves relate to the different policy areas, which are listed in the legend.

[11] In the appendix (p. 11), we report the $\delta_{p(j),s(j)}$ parameters, which show that there is substantial variation in the degree to which respondents think that arguments are persuasive as a function of which side of which issues they are on.

Two main patterns of interest arise in Figure 2. First, there is some variation in the average self-reported persuasive power of our 14 rhetorical elements. The estimates suggest that respondents have a clear aversion to arguments based on ad hominem attacks impugning the characters or motives of those on the opposite side of the issue, as well as to arguments that are based on metaphor and imagery. Previous research has argued that metaphors have large effects on how individuals reason about solutions to social problems like crime (Thibodeau and Boroditsky 2011) and also appear to help individuals develop an understanding of politics and public policy more generally (Bougher 2012; Schlesinger and Lau 2000). Some authors see metaphor as so central to the process of modern political communication that for many politicians, "metaphor is essential to their persuasiveness" (Charteris-Black 2011, 2). Our findings, by contrast, suggest that metaphor-based arguments are less persuasive on average than most other types of rhetorical appeal that we evaluate.

The differences between the other element types are more modest, and it is difficult to be confident about their relative average strength.[12] The posterior probability that the mean strength of arguments based on appeals to authority and expertise is the highest among all element types is 0.52, versus the uniform prior probability of 0.07. We can be reasonably confident that some of the element types are stronger, on average, than others. For example, the posterior probability that appeals to authority are on average more effective is at least 0.9 versus arguments employing appeals to common sense, historical comparisons, populist arguments, appeals to public opinion, metaphors, and ad hominem attacks. Similarly, the probability that populist appeals are, on average, less persuasive than appeals to authority, costs and benefits, side effects, fairness, national greatness, and crisis is approximately 0.9 in each case.[13] Taken together, while the average differences between elements are modest, voters appear to find statements that include references to expertise ("Appeal to authority") and factual argument ("Cost/benefit," "Side effects") more convincing than statements that employ striking language but are thinner in terms of substantive policy-relevant content ("Ad hominem," "Metaphor," "Appeal to populism"). Given that normative concerns about rhetoric center on types of argument that are dedicated "first and foremost to gaining support for a proposition and only secondarily with the merits of the arguments" (Chambers 2009, 337), the ranking we uncover provides a relatively optimistic view of the rhetoric that is deemed persuasive by the UK public.

Second, Figure 2 also clearly illustrates that there is a substantial degree of heterogeneity in the performance of arguments using the same rhetorical element. This is particularly so for certain element types. For example, statements using country comparisons to argue in favor of nationalizing the railways and implementing a sugar tax, and also those arguing against extending the right to buy and reducing tuition fees, are among the most persuasive in our data. By contrast, other arguments of the same type—country comparisons arguing in favor of extending surveillance powers and closing large stores on Boxing Day—are among the weakest that we include in the experiment. Similarly, while appealing to national greatness to oppose fracking is a relatively persuasive way to argue, opposing the expansion of Heathrow using similar appeals is not. Further, argument strength heterogeneity is not equal across all element types. For instance, the "metaphor" arguments tend to perform similarly to one another, and the same is true of "crisis" and "side effect" arguments.

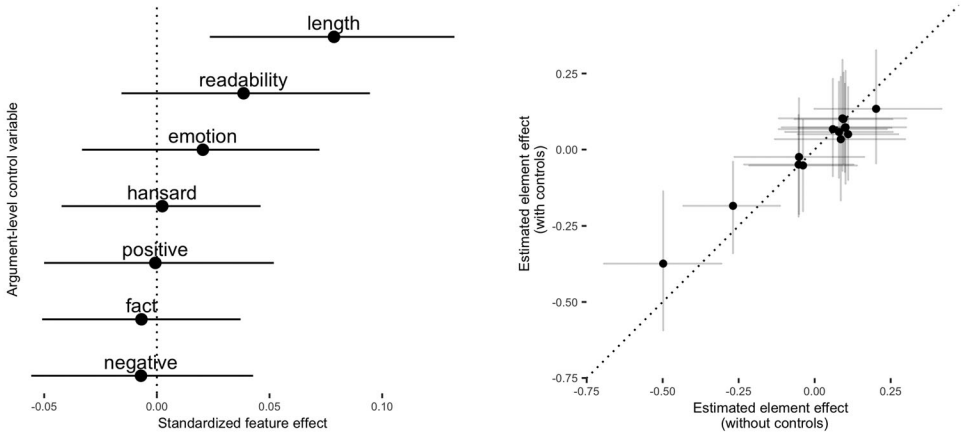
It is important to recognize that these are statements about the treatments that we tested, which may or may not reflect broader populations of arguments that one might define. It might be that we, or the politicians whose statements we adapted, are bad at ad hominem attacks, but that such attacks are effective when deployed more competently. Alternatively, it may be that certain forms of rhetoric—such as the use of metaphor—are less effective in written form than they would be if spoken aloud. Nonetheless, our finding of substantial heterogeneity in the performance of different arguments using the same element type is unlikely to be very sensitive to these concerns. Moreover, all of these criticisms also apply to existing experiments that use single-text implementations of political communication styles. In some contexts, researchers are clear that their interest is in the efficacy of certain rhetorical elements as they pertain to specific policy areas,[14] but authors frequently aim to make more general claims about the persuasiveness of a

[12]The modest average differences between elements do not appear to be due to lack of engagement. First, we find considerable heterogeneity in persuasive power at the *argument* level (see below), indicating that respondents do distinguish between more and less persuasive arguments. Second, although in 28% of comparisons respondents indicated that the two arguments presented were "about the same," very few respondents gave this intermediate response across multiple comparisons. Only 6% of respondents gave three (out of five) "about the same" responses, 5% gave four such responses, and we never observe a respondent giving five such responses. This suggests that the vast majority of respondents were sufficiently attentive to the task that they were able to adjudicate on the persuasiveness of different arguments in most cases.

[13]We report all possible pairwise comparisons in the appendix (p. 12).

[14]See, for example, Barnes and Hicks (2019) and Feldman and Hart (2016).

FIGURE 3 Controlling for Argument-Level Confounders



Note: The figure shows the control variable coefficient estimates (left) and the comparison of element average effects with and without controls (right) from the first experiment and the associated multilevel model.

given rhetorical element on the basis of experiments that provide evidence from only one or a few policy domains. The conclusion we draw from this analysis is that experimental estimates of the effects of rhetorical styles are likely to vary considerably in both sign and magnitude depending on the specific policies to which they relate.

Controlling for Argument-Level Confounders

Implementing multiple texts per latent treatment of interest helps to account for confounding by other text features by allowing us to average over the varying effects of those other features. However, this will only recover unbiased estimates of the latent treatment effects of interest if the variation in text features is uncorrelated with the latent treatments. Our design allows us to further mitigate this problem in cases where we can directly measure the background features that are a cause for concern. Because we have hundreds of treatment implementations, along with a model for the effectiveness of these individual treatments, we can control for potential confounding features when estimating the element effects. We adapt Equation (3) to include a vector of K argument-level measures, which we denote $x_{k,j}$:

$$\begin{aligned} \alpha_j = \delta_{p(j),s(j)} + \mu_{e(j)} + \sum_{k=1}^{K} \gamma_k x_{k,j} + \nu_j \\ \mu_e \sim N(0, \omega) \\ \nu_j \sim N\left(0, \sigma_{e(j)}\right). \end{aligned} \tag{4}$$

The parameters γ_k represent conditional average linear effects of text feature k on argument strength. We have identified seven argument-level variables that represent features of our argument texts that might plausibly confound the effects of the rhetorical elements that we aimed to study: argument length, readability, positive and negative tone, overall emotional language, fact-based language, and whether the argument was based on parliamentary speech from Hansard or was created by the authors of this study.[15]

Figure 3 presents the results. The left-hand panel of the plot shows the standardized posterior point estimates and intervals of the γ_k parameters from Equation (4) and the right-hand panel compares the point estimates of the element average parameters $\mu_{e(j)}$ from the models with (Equation 3) versus without controls (Equation 4).[16]

Of the seven argument-level control variables we include, only length has a clearly significant effect on argument persuasiveness. On average, respondents find arguments with more words somewhat more persuasive than arguments with fewer words. We find weak evidence that readability and emotional content positively predict persuasiveness. There is no difference in average quality between the arguments written by the authors of this study and those from the parliamentary record.

[15]Length is measured as the number of words in the argument, and readability using the Flesch's Reading Ease Score (Flesch 1948). We measure tone using the proportion of words in each argument listed in the positive and negative categories of the Affective Norms for English Words dictionary (Nielsen 2011); emotion using the "affect" category from the 2015 Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker, Francis, and Booth 2001)); and fact-based language using in the "quantitative" and "numeric" categories of the LIWC dictionary.

[16]The dotted line in the right-hand panel of the figure is the 45-degree line.

The right-hand panel of the figure demonstrates that controlling for the additional text features has limited consequences for the estimated rhetorical style effects. There is a slight attenuation of the differences between the rhetorical styles because the two least popular element types had arguments that were somewhat shorter than average. Thus, we find little evidence of what is conventionally called "confounding" (which Grimmer and Fong 2022 call "aliasing") in this context. The differences in the performance of our text treatments based on different rhetorical elements cannot be explained away by these measurable differences in the implementation of those elements in our text treatments. A major strength of our design is that we are able to assess robustness to alternative explanations for differences in the performance of different textual treatments after the experiment is completed.

In the appendix (pp. 13–16), we illustrate alternative multilevel models that we can fit to these data. We show that the strength of arguments is generally positively correlated across respondents of different age, education, attention to politics, and even past vote: Arguments tend to be more or less effective for everyone, with limited heterogeneity across groups. With another variation on this model, we show that the relative strength of arguments is similar when compared to arguments on the same side of the issue as they are when compared to arguments on the other side of the issue. With a third variation, we show that the for and against arguments on the same issue and element have correlated efficacy: Some argument types may be a better match for some issues, regardless of which side the argument is made for.

From Persuasiveness to Persuasion

Our estimates of argument strength are based on responses to questions that prompt individuals to assess which arguments they find to be more persuasive. Do responses to these self-assessment questions (and the modeling approach that we apply to them) in fact identify arguments that, when delivered out of sample, actually persuade new respondents to endorse different policy positions? To answer this question, we fielded a validation experiment with YouGov to new respondents in February 2020—eight months and one general election after the initial experiment—which we use to evaluate whether comparisons of argument persuasiveness translate into arguments that are more effective at changing opinion.

Design

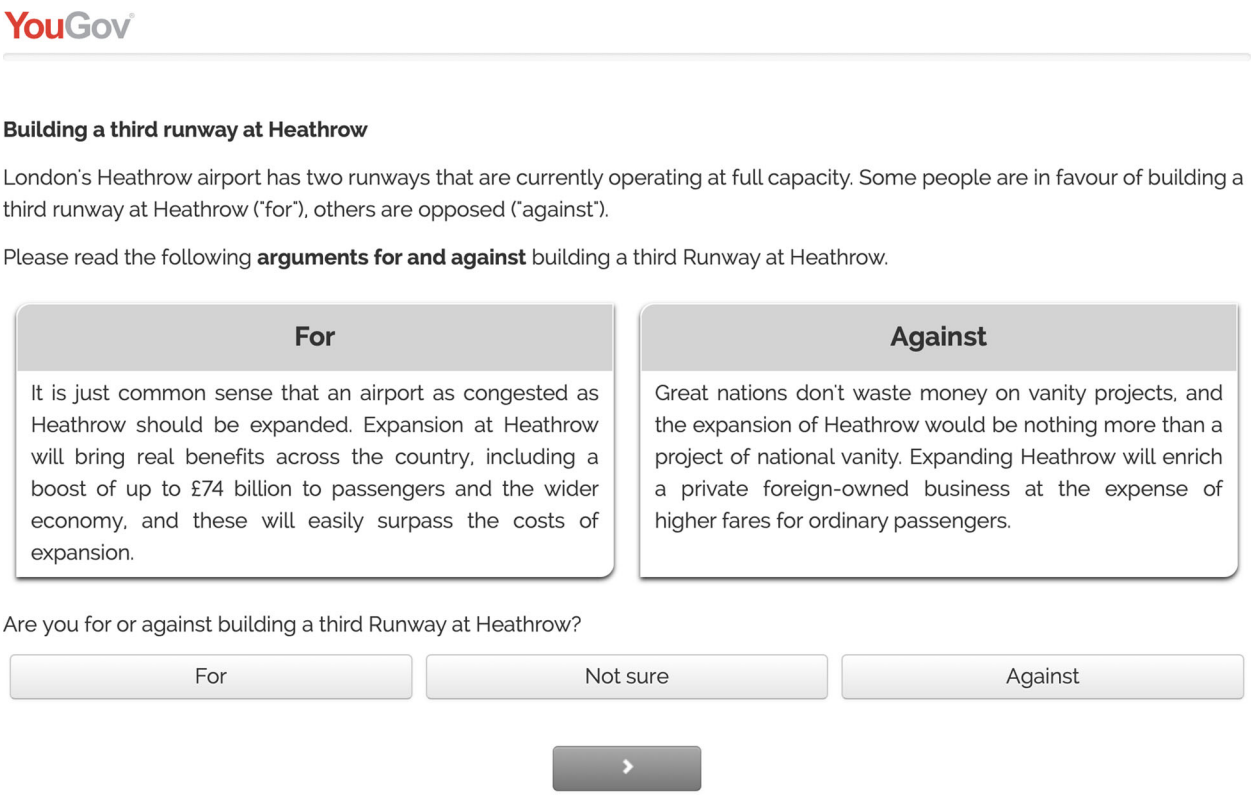
In this validation experiment, our treatments are constructed from the arguments that we used in our initial experiment. Rather than presenting single arguments on each side, we present paragraphs constructed from multiple arguments. This sets a more demanding standard for validation: The differences in argument strength have to be robust to being presented with other arguments, and not merely apply when the argument is presented alone.

Using our estimates of argument strength ($\mu_{e(j)} + \nu_j$), we select the three most and three least persuasive arguments in favor of and against each policy, which we concatenate to form short treatment paragraphs. As shown in Figure 4, each respondent sees two opposing paragraphs: one with arguments in favor of the policy, and one with arguments against the policy. For each policy area, we define two treatment conditions. In the "strong in favor" condition, respondents see the strongest arguments in favor of a policy and the weakest arguments against it (according to the estimates from the initial experiment). In the "strong against" condition, respondents see the weakest arguments in favor of a policy, and the strongest arguments against it. We collected responses on two randomly selected issues from each of our 6,600 respondents, giving us a total of 13,200 responses.

As we show in Figure 3, argument strength is correlated with sentence length. If our treatment paragraphs here always used equal numbers of arguments, then the strong paragraphs would contain more words on average than the weak paragraphs, and we might be concerned that differences in length might confound the effects of our latent quantity of interest, argument strength. Once again applying the idea that having some measurable variation in treatment texts allows us to address potential confounding, we randomly vary whether each paragraph is made up of all three or just two of the three strongest/weakest arguments from our initial experiment. Our analysis presented below averages over this variation. A further multilevel model analysis presented in the appendix (pp. 16–17) confirms that the conclusions hold when we control for the number of arguments and number of words in each paragraph.

A key difference between this validation experiment and our original experiment is that here we do not ask respondents to identify persuasive arguments, but instead directly ask, "Are you for or against <policy issue>?"; respondents can select "For," "Not sure," or "Against." If we have truly identified persuasive sets of arguments, we should see a greater fraction of respondents endorsing the policy among those who see strong arguments in

FIGURE 4 Survey Prompt for Experiment 2



Note: Prompt fielded to 6,600 respondents in February 2020.

favor of that policy, and a smaller fraction endorsing the policy among those who see strong arguments against the policy.

In addition, the estimates from our initial experiment suggest bigger differences in the persuasiveness of argument sets in some policy areas than in others. In some policy areas we observe very strong strong arguments and very weak weak arguments, whereas in other policy areas we observe only moderately strong strong arguments and moderately weak weak arguments. In this validation experiment, when we pair strong arguments and weak arguments in each policy area, we have *ex ante* variation in expected treatment effect sizes: The magnitude of the treatment effects in Experiment 2 should correlate positively with the expected difference in treatment strengths across policy areas as measured from Experiment 1.

To evaluate this expectation, we define the expected difference in argument strength for sets of three arguments on either side of a policy issue as the average of the strengths of the individual arguments:

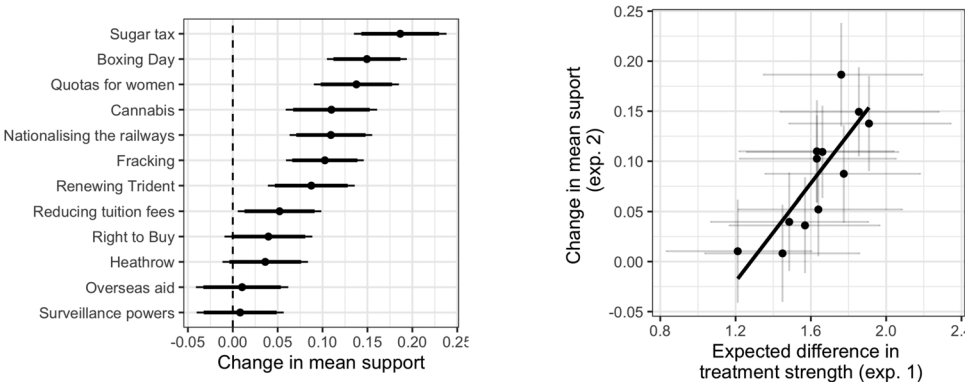
$$\pi = \left(\frac{1}{3}\sum_{j \in \text{in favor}} \mu_{e(j)} + \nu_j\right) - \left(\frac{1}{3}\sum_{j \in \text{against}} \mu_{e(j)} + \nu_j\right). \quad (5)$$

For each policy issue, this calculation yields one relative strength value for the treatment condition where the "in favor" arguments are the strongest in our data and the "against" arguments are the weakest ($\pi_{\text{strong in favor}}$), and one where the "against" arguments are strong and the "in favor" arguments are weak ($\pi_{\text{strong against}}$). The expected strength of the treatment when we compare the "strong in favor" and the "strong against" conditions is therefore given by the difference between these two quantities: $\pi_{\text{strong in favor}} - \pi_{\text{strong against}}$. This quantity will be larger for those policy issues where our initial analysis revealed greater variation in treatment strength across both "in favor" and "against" arguments.

Results

The results are given in Figure 5. There are three main conclusions from this analysis. First, the left-hand panel of Figure 5 shows the simple difference in mean support for each policy between respondents in the "strong in favor" condition and the "strong against" condition. The point estimates of the treatment effects are positive for every policy area, and most of them are significantly different from zero.

FIGURE 5 Comparing Persuasiveness and Persuasion



Note: The left-hand panel gives the difference in mean support between respondents in the "strong in favor" and "strong against" conditions in Experiment 2. The right-hand panel compares the treatment effect estimates from Experiment 2 with the expected differences in argument strength as measured in Experiment 1.

Second, many of the treatment effects we estimate are very large. The largest treatment effect we estimate is for the "Sugar tax" issue, where respondents in our "strong in favor" condition are 19 (95% interval: 13–24) percentage points more likely to endorse the policy than respondents in our "strong against" condition. Similarly, for the "Boxing Day," "Quotas for women," "Cannabis," "Nationalizing the railways," and "Fracking" issues, our point estimates imply that the strong arguments we deploy in favor of those policies persuade more than 10 percentage points of respondents to endorse the policy relative to when we deploy strong arguments against the policy. The large size of these effects suggests that the experimental design and modeling strategy that we describe are successful in measuring the relative persuasive power of different political arguments.

Third, the treatment effects vary considerably in magnitude across policy issues. The right panel of Figure 5 provides very strong evidence that expected differences in argument strength from Experiment 1 predict the magnitude of the treatment effects from Experiment 2. The *y*-axis measures the change in mean support for each policy area from our experiment, and the *x*-axis measures the expected difference in treatment strength based on our estimates from our first experiment (Equation 5). These quantities are clearly positively related across the 12 issue areas, and the linear association between the two is significantly different from zero ($t = 4.53$). Larger interval differences on our argument strength scale measured in Experiment 1 translate into larger persuasion effects when tested out of sample in Experiment 2.

Our measure of the expected difference in treatment strengths implicitly assumes that the effects of arguments as measured in our initial experiment are *additive*. However, respondents may not evaluate combinations of arguments on the basis of their average persuasive appeal. For instance, respondents may be persuaded by the set of arguments that contains the single strongest argument. Our experiment is not designed to test whether alternative argument aggregation rules of this sort would be more predictive of persuasion, but this is an important avenue for future work.

Conclusion

Identifying which general argument forms are most persuasive is a long-standing goal in the study of politics. For Aristotle (2004, 1355b), rhetoric entails "discovering the possible means of persuasion in reference to any subject whatsoever." However, though contemporary empirical work has established the persuasive effects of certain argument types in certain circumstances, scholarship on persuasion has not clearly decomposed the sources of persuasive appeal into distinct rhetorical elements. The central substantive finding of our study suggests why this might be the case: it is very difficult to identify rhetorical strategies that are *consistently* more persuasive than others when considered across multiple policy issues. Basing our design on the types of rhetoric that are regularly found in real-world political speeches in the UK, we found only moderate average differences in the persuasive power of 14 different rhetorical elements, and we demonstrated that there is significant heterogeneity in argument strength within element types. Together, these findings imply that the persuasiveness of different argu-

ment types is likely to be highly context-dependent, and that analyzing the rhetorical structure that characterizes an argument allows us to predict the persuasiveness of that argument only to a limited extent.

Our empirical findings also reinforce two key methodological points. First, they imply external validity concerns for existing studies that rely on single implementations of latent treatments in texts. Though some existing persuasion studies make issue-specific claims, many seek to offer more general conclusions about the effectiveness of different forms of rhetoric. The effect heterogeneity we uncover for arguments of the same rhetorical type suggests that the persuasive power of a given rhetorical element may be very different across issues and therefore that such generalizations should be made with caution. Second, they suggest that researchers should more generally use designs based on pooling evidence from many small implementations rather than a few large ones.

In making this point about the external validity of other studies, we are arguably guilty of the same kind of extrapolation that we are cautioning against. In a narrow sense, we have demonstrated that arguments of the types frequently used in the UK parliament vary widely in their ability to persuade UK citizens, across a set of medium-salience UK political issues. Does this translate to other kinds of survey experiments that political scientists use to assess theories of public opinion and political psychology? We cannot clearly demonstrate that it does. Nonetheless, we think that our empirical results usefully demonstrate a general theoretical concern, which clearly applies *as a concern* across a wide range of studies. Some experimental domains may not exhibit this level of implementation-level heterogeneity, for various reasons. But, at the very least, a strong theoretical argument ought to be expected when researchers move from their specific experiment to more general claims about an underlying phenomenon. Better than such an argument, though, would be more widespread use of the core approach of this article: conducting a larger number of smaller experiments and using multilevel models to characterize the distribution of results.

Our design and results suggest several new avenues of research into the persuasive effects of different rhetorical strategies. First, our experiment employs written texts, but rhetorical skill may manifest differently in spoken and written forms. It is plausible that the ordering of elements that we describe would change if we used videos of politicians speaking rather than texts of their speeches as the basis of our experiment. Using multiple treatment implementations to capture latent treatment effects would apply equally well to video-based as to text-based treatments. Additionally, video treatments would allow researchers to assess a wider variety of rhetorical elements that are difficult to capture adequately in written form.

Second, our analysis has implicitly assumed that the persuasive effects of rhetorical elements are additive, but this may miss important interactions between elements. For instance, observers of contemporary political debates might believe that ad hominem attacks are likely to be more effective when combined with populist appeals of different types. Our design gives a framework for evaluating such hypotheses. Investigating whether and how interactions of this sort affect respondent choices would be informative about how people process rhetorical appeals, and it could help to reveal how arguments could be combined to optimize their persuasive effects.

Third, interactions may also exist between elements and issues. While we have focused on establishing the average persuasiveness of rhetorical elements *across* issues, researchers could use our experimental framework to begin the process of accumulating evidence on the effectiveness of different argument types *within* specific policy domains.

Fourth, our estimates reflect the effects of only short-run exposure to different types of rhetoric. An interesting further development of the findings we present here would be to embed our experimental design in a panel study, which would allow researchers to evaluate how persuasion effects vary as voters are exposed to rhetorical strategies over a longer period of time.

Finally, our model characterizes the strength of a set of rhetorical elements that we defined *ex ante*. Future work might focus on *discovering* sets of textual features that are predictive of persuasiveness, a task that might be better achieved through a different modeling approach. For instance, when researchers have a corpus of preexisting texts, they might characterize those texts with a large number of features, such as the presence and absence of particular words, or measures of textual complexity, or a distribution over topics. In such a case, working out which of the (potentially very many) features are most predictive of persuasion would be the key challenge, and so a more flexible, regularized model—such as ridge or lasso regression—would be an appropriate choice.

References

Arceneaux, Kevin. 2012. "Cognitive Biases and the Strength of Political Arguments." *American Journal of Political Science* 56(2): 271–85.

Aristotle. 2010. *Rhetoric*. Whitefish, MT: Kessinger.

Atkins, Judi, and Alan Finlayson. 2013. "'… A 40-Year-Old Black Man Made the Point to Me': Everyday Knowledge and the Performance of Leadership in Contemporary British Politics." *Political Studies* 61(1): 161–77.

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(2): 226–42.

Barnes, Lucy, and Timothy Hicks. 2021. "Are policy analogies persuasive? The household budget analogy and public support for austerity." *British Journal of Political Science*, FirstView. 1–19.

Bartels, Larry M. 2003. "Democracy with Attitudes." In *Electoral Democracy*, ed. Michael MacKuen and George Rabinowitz. Ann Arbor: University of Michigan Press, 48–82.

Bechtel, Michael M., Jens Hainmueller, Dominik Hangartner, and Marc Helbling. 2015. "Reality Bites: The Limits of Framing Effects for Salient and Contested Policy Issues." *Political Science Research and Methods* 3(3): 683–95.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351–68.

Bos, Linda, Wouter van der Brug, and Claes H. de Vreese. 2013. "An Experimental Test of the Impact of Style and Rhetoric on the Perception of Right-Wing Populist and Mainstream Party Leaders." *Acta Politica* 48(2): 192–208.

Boudreau, Cheryl, and Scott A. MacKenzie. 2014. "Informing the Electorate? How Party Cues and Policy Information Affect Public Opinion about Initiatives." *American Journal of Political Science* 58(1): 48–62.

Bougher, Lori D. 2012. "The Case for Metaphor in Political Reasoning and Cognition." *Political Psychology* 33(1): 145–63.

Bradley, Ralph Allan, and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3/4): 324–45.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1): 1–32.

Chambers, Simone. 2009. "Rhetoric and the Public Sphere: Has Deliberative Democracy Abandoned Mass Democracy?" *Political Theory* 37(3): 323–50.

Charteris-Black, Jonathan. 2011. *Politicians and Rhetoric: The Persuasive Power of Metaphor*. Palgrave Macmillan. Basingstoke, UK.

Chong, Dennis, and James N. Druckman. 2007a. "A Theory of Framing and Opinion Formation in Competitive Elite Environments." *Journal of Communication* 57(1): 99–118.

. 2007b. "Framing Public Opinion in Competitive Democracies." *American Political Science Review* 101(4): 637–55.

. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104(4): 663–80.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3): 613–28.

Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2019. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6: 1–6.

Dewan, Torun, Macartan Humphreys, and Daniel Rubenson. 2014. "The Elements of Political Persuasion: Content, Charisma and Cue." *The Economic Journal* 124(574): F257–92.

Disch, Lisa. 2011. "Toward a Mobilization Conception of Democratic Representation." *American Political Science Review* 105(1): 100–114.

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98(4): 671–86.

Dryzek, John S. 2010. "Rhetoric in Democracy: A Systemic Appreciation." *Political Theory* 38(3): 319–39.

Elster, Jon. 1998. "Deliberation and Constitution Making." In *Deliberative Democracy*, ed. Jon Elster. New York: Cambridge University Press, 97–122.

Feldman, Lauren, and P. Sol Hart. 2016. "Using Political Efficacy Messages to Increase Climate Activism: The Mediating Role of Emotions." *Science Communication* 38(1): 99–127.

Finlayson, Alan. 2007. "From Beliefs to Arguments: Interpretive Methodology and Rhetorical Political Analysis." *British Journal of Politics and International Relations* 9(4): 545–63.

Flesch, Rudolph. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32(3): 221–33.

Graham, Matthew, and Alexander Coppock. 2021. "Asking about Attitude Change." *Public Opinion Quarterly* 85(1): 28–53.

Grimmer, Justin, and Christian Fong. 2022. "Causal Inference with Latent Treatments." *American Journal of Political Science*, EarlyView.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112(8): 2395–2400.

Hameleers, Michael, Linda Bos, and Claes H. de Vreese. 2017. "'They Did It': The Effects of Emotionalized Blame Attribution in Populist Communication." *Communication Research* 44(6): 870–900.

Hameleers, Michael, and Desirée Schmuck. 2017. "It's Us against Them: A Comparative Experiment on the Effects of Populist Messages Communicated via Social Media." *Information, Communication & Society* 20(9): 1425–44.

Hopkins, Daniel J., and Jonathan Mummolo. 2017. "Assessing the Breadth of Framing Effects." *Quarterly Journal of Political Science* 12(1): 37–57.

Jerit, Jennifer. 2009. "How Predictive Appeals Affect Policy Opinions." *American Journal of Political Science* 53(2): 411–26.

Jung, Jae-Hee. 2020. "The Mobilizing Effect of Parties' Moral Rhetoric." *American Journal of Political Science* 64(2): 341–355

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112(1): 148–66.

Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics* 69(4): 1176–1209.

Lecheler, Sophie, Claes de Vreese, and Rune Slothuus. 2009. "Issue Importance as a Moderator of Framing Effects." *Communication Research* 36(3): 400–425.

Leeper, Thomas J., and Rune Slothuus. 2018. "Can Citizens Be Framed? How Information, Not Emphasis, Changes Opinions." *Working Paper*.

Loewen, Peter John, Daniel Rubenson, and Arthur Spirling. 2012. "Testing the Power of Arguments in Referendums: A Bradley–Terry Approach." *Electoral Studies* 31(1): 212–21.

Mance, Henry. 2016. "Britain Has Had Enough of Experts, Says Gove." *Financial Times*, June 3, 2016.

Nelson, Thomas E. 2004. "Policy Goals, Public Rhetoric, and Political Attitudes." *Journal of Politics* 66(2): 581–605.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(3): 567–83.

Nielsen, Finn Årup. 2011. "A New Anew: Evaluation of a Word List for Sentiment Analysis in Microblogs." arXiv Preprint arXiv:1103.2903.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. "*Linguistic Inquiry and Word Count: LIWC 2001*." Mahwah, NJ: Erlbaum.

Rhetorica ad herennium. 2022. *Rhetorica ad Herennium*.

Riker, William H. 1990. "Heresthetic and Rhetoric in the Spatial Model." In *Advances in the Spatial Theory of Voting*, ed. James M. Enelow and Melvin J. Hinich. New York: Cambridge University Press, 46–50.

Schlesinger, Mark, and Richard R. Lau. 2000. "The Meaning and Measure of Policy Metaphors." *American Political Science Review* 94(3): 611–26.

Sniderman, Paul M., and Sean M. Theriault. 2004. "The Structure of Political Argument and the Logic of Issue Framing." In *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*. Willem E. Saris and Paul M. Sniderman (eds), Princeton University Press: Princeton, New Jersey, 133–65.

Thibodeau, Paul H., and Lera Boroditsky. 2011. "Metaphors We Think With: The Role of Metaphor in Reasoning." *Plos One* 6(2): e16782.

Vavreck, Lynn. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* 2(4): 325–43.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Response Data by Element and Policy
Appendix B: Sensitivity to modelling assumptions
Appendix C: Additional Model Parameter Estimates
Appendix D: Differential Persuasiveness by Respondent Characteristics
Appendix E: Same-side vs Opposite-side Argument Comparisons
Appendix F: Rhetorical Fit to Specific Issues
Appendix G: Multilevel Model for Validation Experiment
Appendix H: MCMC diagnostics
Appendix I: Arguments