The Variable Persuasiveness of Political Rhetoric *

Jack Blumenau *University College London*
Benjamin E Lauderdale *University College London*

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.

Word count: 9999

*__This version__: November 06, 2020

Introduction

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; emphasise a moral rationale for supporting the policy; 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 sustained over the course of millennia (Aristotle, c.322 BCE; Rhetorica ad herennium, c.80 BCE; Riker, 1990; Charteris-Black, 2011), and more recent work has begun to test the efficacy of different communication strategies (Loewen, Rubenson and Spirling, 2012; Thibodeau and Boroditsky, 2011; Schlesinger and Lau, 2000; Bougher, 2012; Lau, Sigelman and Rovner, 2007; Bos, Van Der Brug and De Vreese, 2013; Hameleers, Bos and de Vreese, 2017; Hameleers and Schmuck, 2017; Jung, Forthcoming; Nelson, 2004; Boudreau and MacKenzie, 2014; Jerit, 2009), 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 non-zero effectiveness of specific types of arguments 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 prioritises emotion and passion over reason, and inhibits rational deliberation between citizens (Elster, 1998). However, recent

work in normative political theory which 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 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.

In this paper, 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 modelling approach. We examine types of arguments frequently made in contemporary British politics, and especially in speeches delivered by politicians in the UK parliament. As we discuss below, 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 UK. 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 these forced-choice comparisons 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 effectiveness at

persuading democratic citizens, this paper 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 towards 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, 2019) 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 recognise 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 modelling to estimate this distribution using a large number of implementations, each of which would be statistically underpowered if analysed 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 post-experimental 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 and authority 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 relevant figures of 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. While 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*.

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 which relate to single policy issues. While it is not novel to observe that the effects of particular experimental implementations may not generalise to other domains, we directly quantify the substantial

[1] Britain has had enough of experts, says Gove, Financial Times, 3 June, 2016

variance of the effects of the same treatment types across issues.

Ultimately, our goal is to understand which types of arguments induce voters to support or oppose policy proposals on different issues. However, we note that *persuasion* of this sort is different from the judgements of argument *persuasiveness* that we elicit in our experiment. Survey respondents are likely to overstate changes when they are asked to give self-assessments of the effects of an experimental treatment on their political attitudes (Vavreck et al., 2007; Graham and Coppock, 2019), and responses to our experiment may also be subject to social desirability biases, as respondents might state preferences for arguments they think they *ought* to find more persuasive.

To address these concerns, 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 generalising 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 which are thought to be one source of their persuasive appeal, and we share the understanding of Atkins and Fin-

layson (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 (eg Aristotle, c.322 BCE; 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, which may feature in many different arguments, 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, Forthcoming; Nelson, 2004); appeals based on the expected costs and benefits of policy (Jerit, 2009; Riker, 1990); and the use of expert cues and endorsements (Boudreau and MacKenzie, 2014; Dewan, Humphreys and Rubenson, 2014; Atkins and Finlayson, 2013).

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 emphasise 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 oth-

ers. 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: a tendency to look at different rhetorical elements individually rather than comparatively; a lack of attention to heterogeneity in the effectiveness of rhetorical elements across different issues; and a lack of evidence to support inferences about the effects of "latent" rhetorical types from evidence about specific implementations of those types. We discuss each of these in turn.

First, the cumulative evidence from these studies suggest that elite communication can substantially shift public opinion, and several authors express concern that such results suggest that citizens do not hold stable and well-formed preferences (Druckman, 2004; Disch, 2011). Others have argued that there is an important role for rhetoric in the process of democratic deliberation (Dryzek, 2010), and that certain *forms* of rhetoric are more defensible than others (Chambers, 2009; Kock, 2007). Chambers (2009, 328), for example, emphasises 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 more which we have not cited) the persuasiveness of the relevant style or frame of interest – populism; metaphor; morality;

etc – 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, 2007, 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 which 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 generalise 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."

Some existing research (Lecheler, de Vreese and Slothuus, 2009; Hopkins and Mummolo, 2017) suggests that the persuasive effects of different frames vary across policy issues and it seems plausible *a prior* that certain types of rhetoric may be more appropriate for certain policies 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

[2] See, for example, Bos, Van Der Brug and De Vreese (2013), Arceneaux (2012), Jung (Forthcoming), Nelson (2004), Jerit (2009).

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)? Despite the fact that single-context studies are not informative in this regard, the desire of scholars to generalise from treatment effects that relate to specific policy contexts to broader conclusions about the effectiveness of different rhetorical types is often evident in how authors discuss their findings. Our second contribution, therefore, is to provide evidence of the distribution of effectiveness of rhetorical elements across a wide range of political issues.

Our approach also helps us to overcome a third methodological problem that is common to vignette-style experiments which use single-text treatments. Grimmer and Fong (2019) argue that latent treatments which 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.

Providing several texts per latent treatment allows background features which might confound the latent concept 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, the fact that we have a large number of 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 (rhetorical elements) that motivate our study.

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 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", "Nationalisation 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 emphasise the desirability or undesirability of a given policy. We based our categorisation of rhetorical elements on close reading of contemporary political debates. We began with a short list of possible rhetorical categories, and then expanded and refined our categorisation 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 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 full set of rhetorical elements that we consider, which was not

intended to be exhaustive, is given in table 1.

While our primary goal is to quantify the persuasiveness rhetorical appeals used in contemporary politics, our design is amenable to any arbitrary categorisation of arguments into types so long as the researcher is able to write multiple implementations of the same treatment concepts. An alternative approach, for example, would be to use a pre-existing categorisation of either contemporary (eg Charteris-Black, 2011) or classical (eg Aristotle, c.322 BCE) rhetorical appeals. Similarly, we also considered other rhetorical elements that feature in UK debate that we might have included – such as "examples of personal narrative" and "appeals to freedom" – but did not, as we felt that it would be too difficult to write treatments across all the issues we included in our sample. In general, however, our design could be used to investigate the relative persuasiveness of a wide range of alternative rhetorical categories.

An **argument** is a text that makes a case in favour 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. For each policy issue, we wrote two separate arguments for each of the rhetorical elements: one arguing in favour of the policy, and one arguing against. This results in 14 x 12 x 2 = 336 separate arguments which are the basic treatments in our experiment.

To ensure the arguments we used resembled the types of argument used by politicians in the UK, 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 edited these texts into the form we use in the experiment. In the appendix we present all 336 arguments we include, and for many of the sentences we provide hyperlinks to the source documents on which our treatments are based. When it was not possible to identify an

Table 1: Elements

Element	Description
Appeal to fairness	A statement based on appeals to fairness. Uses the root "fair".
Costs vs benefits arguments	A statement which makes an explicit argument based on the costs and/or benefits of a policy. Uses the root “cost” and/or “benefit.
Country comparison	A statement is made about this policy or a similar policy in a named country, set of countries, or uses language about generic countries. This may be a statement of fact about whether the policy exists, or may be making an argument about its success/failure.
Crisis	A statement which emphasises the attractiveness or unattractiveness of a policy based on an argument that something is or is not a crisis. Must include the word "crisis".
Side-effects	A statement which emphasises the side-effects of a policy in order to persuade. Includes the phrase "unintended consequence/effect” or "side effect".
Metaphor/figure of speech	A statement which uses a figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable for rhetorical or vivid effect. May be hyperbolic.
Ad hominem	A statement which makes appeals based on undermining or impugning the motives of those on the other side of the argument. Might include mentions of corruption, ulterior motives, biased agendas, lack of consideration, hypocrisy, bad faith.
Appeals to expertise	A statement which reports the view of an entity with relevant subject area expertise in support of an argument. Explicitly mentions a not explicitly partisan entity – such as a professional body, academic organisation, research institute, think tank, union, business group, etc – by name.
Appeal to history	A statement of evidence from past policy experience *in the UK*. Includes explicit references to certain years and/or historical periods, or uses generic language about "the past" or "in previous years" or past generations.
Appeal to national greatness	A statement based on an appeal to national pride. Uses language about the UK being a world-leading country in this policy area, uses the word "great" as a descriptive, and/or makes explicit appeals to British values. Mentions the phrases "Britain" or "the UK" or "British" or "this country".
Appeal to populism	A statement which makes distinctions between elites and non-elites as the basis of a rhetorical appeal for the policy, situating the argument on the side of the non-elites. Does not require specific language to identify elites / non-elites, but can use familiar stand-ins to represent these categories.
Common sense	A statement which argues for or against a policy based on appeals to common sense or reasonableness. States that an argument for/against the policy is "common sense".
Morality	A statement which makes arguments for or against a policy based on morality concerns. Includes specific mentions of things being moral/immoral or right/wrong.
Public opinion	A statement which bases its argument on a claim about public opinion. Includes a phrase which has some quantifier (not necessarily numeric) about the support or opposition of the public for a policy.

example of our rhetorical styles in the texts of the Commons' debates on a particular issue, we wrote arguments of our own, making the texts as similar as possible in style to those based on politicians' speeches.[3]

Survey instrument

We use these arguments as the basis of a forced-choice experiment which was fielded by YouGov to their UK online panel in June 2019. Following an introduction screen describing the task, respondents were randomly presented with two arguments pertaining to a particular policy issue and asked which of the two arguments they found more persuasive. Policy issues were sampled from the full set of 12 policy issues. For each selected policy, we then randomly sampled whether a respondent was presented with two arguments "in favour" of that policy, two arguments "against" that policy, or one argument "in favour" and one argument "against". In 50% of cases, respondents were presented with two arguments on the same side of a policy, and in 50% of cases they were presented with arguments from alternative sides.[4] We collected responses on four randomly selected issues from each of 3317 respondents, giving us a total of 13268 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, here 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 behaviour and attitudes (Vavreck et al., 2007; Graham and

[3]Below, we control for whether an argument is written by an MP, or by the authors of this study. Somewhat disappointingly, we were no better (or worse) than UK politicians at writing persuasive political arguments.

[4]In the appendix, we report the results of an analysis which examines the degree to which "same side" and "different side" comparisons result in similar rankings of arguments. We find that these different types of comparisons result in rankings that correlate at 0.81, which indicates that we can get nearly the same information from same-side as from opposite-side comparisons.

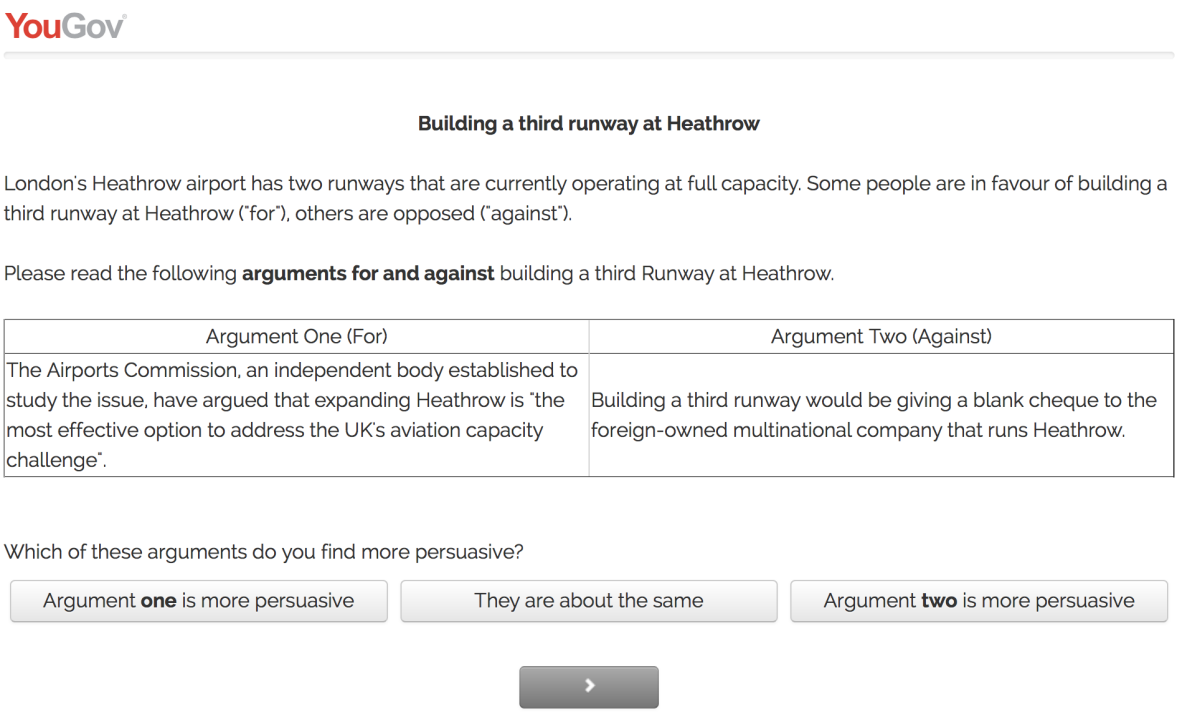

Figure 1: Experiment prompt

Coppock, 2019), and 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 later) 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.

The Relative Persuasiveness of Rhetorical Elements

Modelling Persuasiveness

Our design generates a set of responses that specify "winners" from a pairwise competition between arguments, with the possibility of ties. 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 modelling task is to infer the efficacy of particular types of arguments given the results of the pairwise contests.

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

$$Y_i \in \begin{cases} 1 = \text{Argument 2 is more persuasive} \\ 2 = \text{About the same} \\ 3 = \text{Argument 1 is more persuasive} \end{cases} \tag{1}$$

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:

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

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 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.[5] 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 modelling how the strength of arguments vary as a function of rhetorical features that appear in those arguments. We therefore specify a hierarchical model for the α_j parameters.[6]

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

[6] The hierarchical model described by equation 3 distinguishes our approach from work on Canadian referendum arguments by Loewen, Rubenson and Spirling (2012), who ask survey respondents to make forced-choice comparisons between pairs of political arguments and use the responses to estimate of

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:

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

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 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 that 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 $\mu_{e(j)}$ capture the average effect of each of our rhetorical elements. The final set of parameters are the ν_j, which are argument-specific "residuals" that characterise the distribution of argument-level effects around the element-type average. We estimate separate variance parameters for each element ($\sigma_{e(j)}$).[7] We estimate the model using Stan (Carpenter et al., 2017).[8]

the persuasive power of those arguments via extensions to a Bradley-Terry model. Our project differs in that they are concerned with estimating the relative persuasive power of individual arguments (the αs) whereas we are interested in the power of different rhetorical element types which can appear in many different arguments (the hyperparameters describing the distribution of αs).

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

[8] We used three chains of 500 iterations, after 200 iteration burn-in.

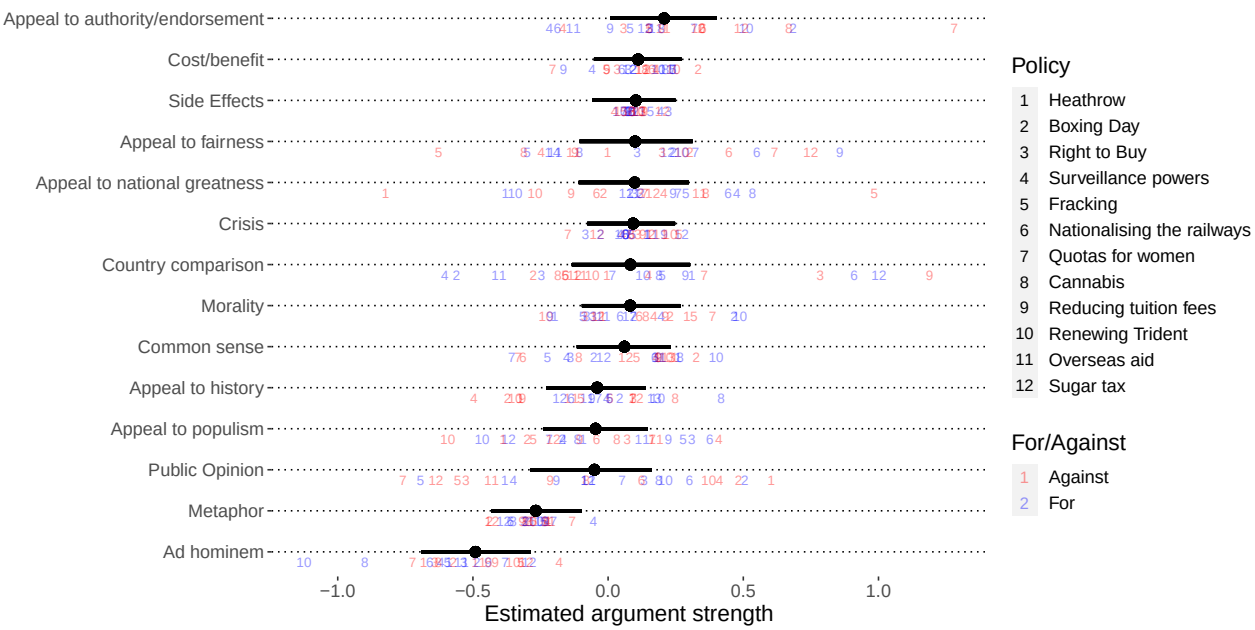

Figure 2: Estimated argument strengths

Results

We present the main results from our model in Figure 2.[9] 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. Blue numbers indicate arguments on the "for" side of the relevant issue, and red numbers indicate arguments on the "against" side of the issue. The numbers themselves relate to the different policy areas, which are listed in the legend.

Two main patterns of interest arise in Figure 2. First, there is some variation in the average persuasive power of our 14 rhetorical elements. The estimates suggest that respondents have a clear aversion to the arguments we provided that were 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. The relatively poor performance of arguments based on metaphor contrasts with findings in

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

the existing literature. 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 in developing understanding of politics and public policy more generally (Schlesinger and Lau, 2000; Bougher, 2012). 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. 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 of the kind that we tested 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 to appeals to authority, costs and benefits, side effects, fairness, national greatness, and crisis is approximately 0.9 in each case.[10]

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

[10]We report all possible pairwise comparisons in the appendix.

to gaining support for a proposition and only secondarily with the merits of the arguments" (Chambers, 2009), 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 favour of nationalising 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 favour of extending surveillance powers and closing large stores on Boxing Day – are amongst 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 recognise 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 MPs 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 very 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,[11] but authors frequently aim to make more general claims about the persuasiveness of a 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 treatment of interest helps to account for confounding by other correlated text features by allowing us to average over those features. However, this will only recover unbiased estimates of the causal effects of interest if the variation in background features is uncorrelated with our latent treatments. Our design allows us to 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, and a model for the effectiveness of these individual treatments, we can simply control for potential confounding features when estimating the element effects. In the following analysis, we expand the second stage of our model to incorporate measurable features of our arguments that may correlate with the rhetorical styles. We adapt equation 3 to include a vector of K argument-level measures, which we denote $x_{k,j}$:

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

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

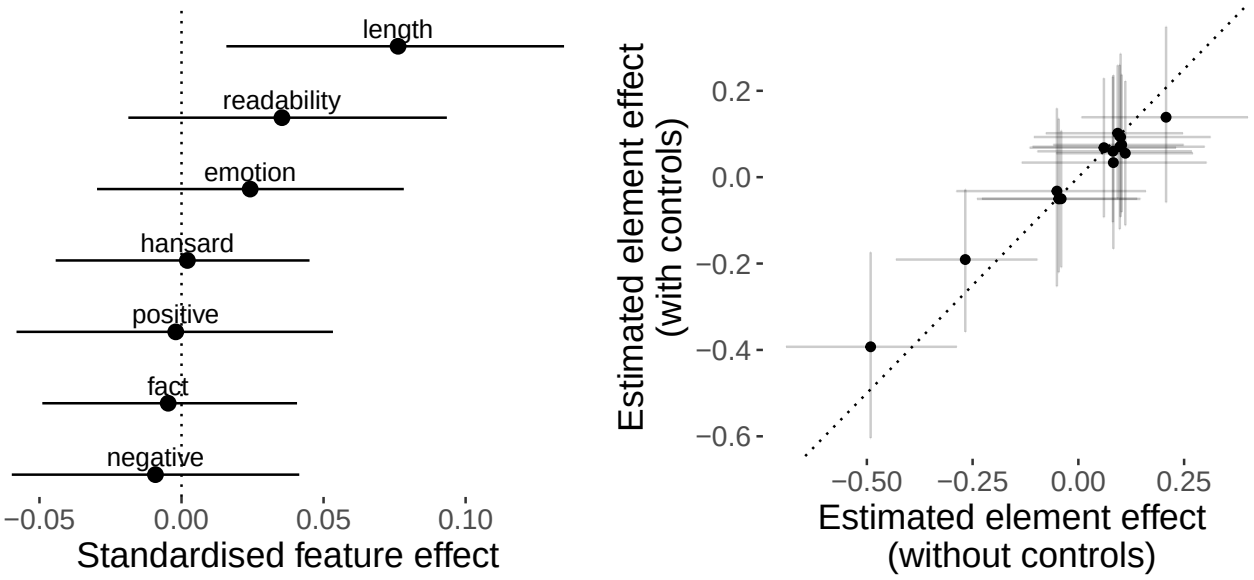

Figure 3: Control variable coefficient estimates (left) and comparison of element average effects with and without controls (right)

The parameters γ_k represent conditional average linear effects of text-feature k on argument strength. We have identified seven argument-level variables which, in each case, represent features of our argument texts that might plausibly confound the effects of our rhetorical styles. We include 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.[12]

Figure 3 presents the results. The left-hand panel of the plot shows the standardised 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).

Of the 7 argument-level control variables we include, only length has a clearly signif-

[12] 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.

icant effect on argument persuasiveness. The estimates in the left-hand panel suggest that, 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.

The right-hand panel of the figure also 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 confounding here. Nevertheless, it is a major strength of our design that we are able to assess robustness to potentially confounding features of this kind, after the experiment is completed.

In the appendix, 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/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 pro and con arguments on the same issue-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 which prompt individuals to assess which arguments they find to be more persuasive. Do responses to these self-assessment questions, and the modelling 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? Is the variation in argument quality that we identify consequential for persuasion? 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 our validation, our treatments are constructed from the arguments that we used in our initial experiment. Using our estimates of argument strength ($\mu_{e(j)} + \nu_j$) we select the most and least persuasive arguments in favour 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 favour of the policy, and one with arguments against the policy. For each policy area, we define two treatment conditions. In the "strong in favour" condition, respondents see the strongest arguments in favour of a policy and the weakest arguments against (according to the estimates from the initial experiment). In the "strong against" condition, respondents see the weakest arguments in favour of a policy, and the strongest arguments against. We collected responses on two randomly selected issues from each of our 6600 respondents, giving us a total of 13200 responses.

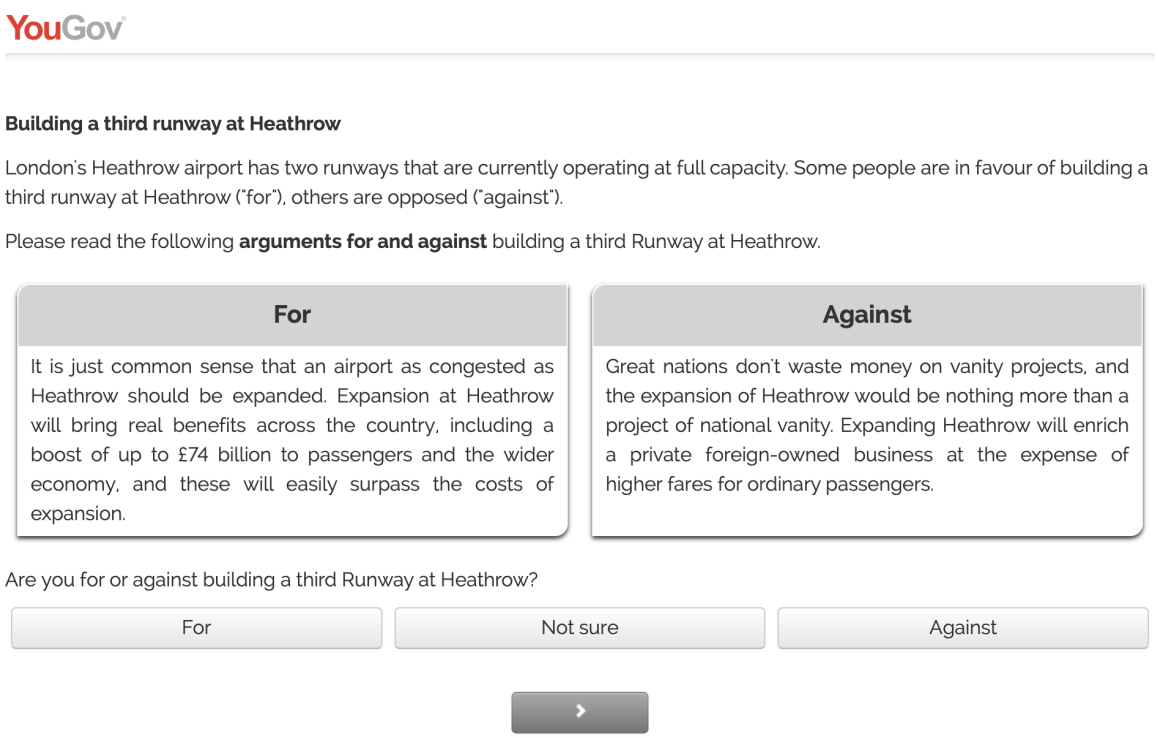

Figure 4: Experiment 2 prompt

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 two or three of the three strongest/weakest arguments from our initial experiment. We then control for the number of arguments and number of words in each paragraph in the model we describe below.

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>?", and 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 favour of that policy, and a smaller fraction endorsing the policy among those who see strong arguments against the policy.

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, while 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 two should correlate positively with the expected difference in treatment strengths across policy areas as measured from experiment one.

Modeling Persuasion

The simplest way to analyse the data from this experiment is to look at differences in means for each issue, where responses endorsing "For" are coded $Y = 1$, "Not sure" $Y = 0.5$, and "Against" $Y = 0$. However, if we want to characterise the set of treatment effects, it is again natural to use a multilevel model. This model takes the form:

$$
\begin{aligned}
Y_{j,j'} &= \theta_p + \alpha_j - \alpha'_j + \epsilon \\
\alpha_j &= \delta_{p(j)} * \text{Strong}_j + \\
&\quad \gamma_{Words} * \text{\# Words}_j + \\
&\quad \gamma_{Arguments} * \text{\# Arguments}_j + \nu_j \\
\nu_j &\sim N(0, \sigma) \\
\delta_p &\sim N(\mu_\delta, \sigma_\delta)
\end{aligned}
$$

where θ_p is the baseline popularity of the "For" side of each policy area p, $\delta_{p(j)}$ is a parameter capturing the effect of using strong (as opposed to weak) arguments on persuasive power for policy area p, γ_{Words} is the effect of the number of words, and $\gamma_{Arguments}$ captures whether paragraphs with three rather than two arguments are more persuasive. In this analysis, our main quantities of interest are $\delta_{p(j)}$—which measure the average treatment effect of going from weak arguments in favour and strong arguments against to strong arguments in favour and weak arguments against for a particular issue area—and μ_δ—which measures the average of these average treatment effects across the set of issues in the experiment. The way that the $\delta_{p(j)}$ and μ_δ are defined in the multilevel model as moving from a weak to a strong argument on one side means that they correspond to half of the experimental difference in means, which correspond to moving from weak to strong on one side and strong to weak on the other.

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 favour" 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 by conventional standards.

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 favour" 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", "Nationalising the railways", and "Fracking" issues, our point estimates imply that the strong arguments we deploy in favour of those policies persuade more than ten percentage points of respondents to endorse the policy

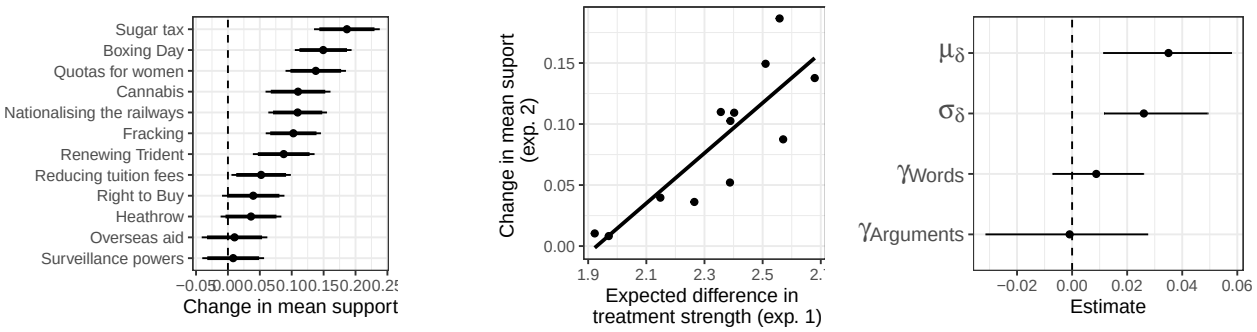

Figure 5: Experiment two results

relative to when we deploy strong arguments against the policy. The large size of these effects suggest that the experimental design and modelling strategy that we describe is successful in measuring the relative persuasive power of different political arguments.

Third, the treatment effects vary considerably in magnitude across policy issues. The middle panel of figure 5 provides very strong evidence that expected differences in argument strength from experiment one predict the magnitude of the treatment effects from experiment two. 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.[13] These quantities are clearly positively related across the 12 issue areas, and the linear association between the two is significantly different from zero ($t = 5.03$). Larger interval differences on our argument strength scale measured in experiment one translate into larger persuasion effects when tested out of sample in experiment two.

The right-hand panel of figure 5 presents median posterior estimates and 95% inter-

[13]We define the expected difference in argument strength for sets of three arguments on either side of a policy issue as:

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

We do this twice for each policy issues, once where the "in favour" arguments are the strongest in our data and the "against" arguments are the weakest ($\pi_{\text{strong in favour}}$), and once where the "against" arguments are strong and the "in favour" arguments are weak ($\pi_{\text{strong against}}$). The expected strength of the treatment in experiment two is therefore given by the difference between these two quantities: $\pi_{\text{strong in favour}} - \pi_{\text{strong against}}$

vals for the key parameters from the multilevel model described in equation 5. Neither the number of words nor the number of arguments presented significantly predict responses. We can rule out the possibility that there is no variation in the average treatment effects across issues (the interval estimate for σ_δ starts well above 0). Finally, our estimate of μ_δ, the average effect of going from a weak to a strong argument, is 0.035, which translates to an average treatment effect of 7 percentage points on the scale of the simple difference in means discussed previously.

Conclusion

We have described a new experimental design and modelling strategy for testing the relative persuasiveness of different types of political arguments. Basing our design on the types of rhetoric that are regularly found in real-world political speeches in the UK, we implemented an experiment using 336 individual arguments pertaining to 14 rhetorical elements across 12 policy issues. Combining a Bradley-Terry style model with a series of hierarchical models, we have demonstrated that there are differences in the persuasive power of different rhetorical elements, with "appeals to authority" among the strongest types we test, and "ad hominem" and "metaphor" among the weakest. However, we also demonstrated that there is significant heterogeneity in argument strength *within* element types, implying external validity concerns for existing studies that rely on single implementations of latent treatments in texts. These empirical findings reinforce the methodological point that researchers should more generally use designs based on pooling evidence from many small implementations rather than few large ones.

Astute readers will note that, 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, one from a domain-specific demonstration to a far more general claim. 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. It may be the case that some experimental domains do not exhibit this level of implementation-level heterogeneity, for various reasons. But at the very least, a very 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 paper: conducting a larger number of smaller experiments, and using multilevel models to characterise the distribution of results. As Grimmer and Fong (2019) note, the confounding issues that motivate our design are not a concern for A/B testing where researchers are interested in evaluating responses to two or more specific texts. However, our approach is appropriate wherever the purpose of an experiment is to illustrate a more general phenomenon, rather than to study the specific treatment(s) being applied.

There are several dimensions on which our study is limited in scope, many of which suggest avenues for future research. First, we use written texts to implement a set of rhetorical elements, 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. Ad hominem attacks may, for example, seem more persuasive when delivered aloud than when read on the page. In addition, using video treatments would also mean that we could assess a wider variety of rhetorical elements, such as the emotional intensity with which arguments are conveyed, which are difficult to capture

adequately in written form. Second, we focus on a set of moderate salience issues in UK politics, but one might imagine that some of our elements would be more or less persuasive on a different set of issues. For example, we might expect arguments that employ expert endorsements or cost/benefit analyses to be less persuasive on issues where respondents already have strongly-held views. Finally, 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. We leave these endeavours for future work.

References

Arceneaux, Kevin. 2012. "Cognitive biases and the strength of political arguments." *American Journal of Political Science* 56(2):271–285.

Aristotle, ([undated]. c.322 BCE. *Rhetoric*. Kessinger Publishing, 2004.

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):161–177.

Barabas, Jason and Jennifer Jerit. 2010. "Are survey experiments externally valid?" *American Political Science Review* 104(2):226–242.

Barnes, Lucy and Timothy Hicks. 2019. "The Household Finance Analogy Does Not Drive Mass Support for Austerity." *Working Paper* .

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):683–695.

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):351–368.

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):145–163.

Bradley, Ralph Allan and Milton E Terry. 1952. "Rank analysis of incomplete block designs: I. The method of paired comparisons." *Biometrika* 39(3/4):324–345.

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).

Chambers, Simone. 2009. "Rhetoric and the public sphere: Has deliberative democracy abandoned mass democracy?" *Political theory* 37(3):323–350.

Charteris-Black, Jonathan. 2011. *Politicians and rhetoric: The persuasive power of metaphor*. Springer.

Chong, Dennis and James N Druckman. 2007. "Framing public opinion in competitive democracies." *American Political Science Review* 101(4):637–655.

Chong, Dennis and James N Druckman. 2010. "Dynamic public opinion: Communication effects over time." *American Political Science Review* 104(4):663–680.

Coppock, Alexander. 2019. "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach." *Political Science Research and Methods* 7(3):613–628.

Coppock, Alexander, Seth J Hill and Lynn Vavreck. 2019. "Persuasive Effects of Presidential Campaign Advertising: Results of 53 Real-time Experiments in 2016.".

Dewan, Torun, Macartan Humphreys and Daniel Rubenson. 2014. "The elements of political persuasion: Content, charisma and cue." *The Economic Journal* 124(574):F257–F292.

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):671–686.

Dryzek, John S. 2010. "Rhetoric in democracy: A systemic appreciation." *Political theory* 38(3):319–339.

Elster, Jon. 1998. "Deliberation and constitution making." *Deliberative democracy* 97:111.

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." *The British Journal of Politics and International Relations* 9(4):545–563.

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of applied psychology* 32(3):221.

Graham, Matthew and Alexander Coppock. 2019. "Asking About Attitude Change." *Working Paper* .

Grimmer, Justin and Christian Fong. 2019. "Causal Inference with Latent Treatments." *Working Paper* .

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):1425–1444.

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.

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):411–426.

Jung, Jae-Hee. Forthcoming. "The Mobilizing Effect of Parties' Moral Rhetoric." *American Journal of Political Science* .

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):148–166.

Kock, Christian. 2007. "Norms of legitimate dissensus." *Informal Logic* 27(2):179–196.

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):212–221.

Nelson, Thomas E. 2004. "Policy goals, public rhetoric, and political attitudes." *The 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):567–583.

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." *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Rhetorica ad herennium. c.80 BCE. *Rhetorica ad herennium*.

Riker, William H. 1990. "Heresthetic and rhetoric in the spatial model." *Advances in the spatial theory of voting* 46:50.

Schlesinger, Mark and Richard R Lau. 2000. "The meaning and measure of policy metaphors." *American Political Science Review* 94(3):611–626.

Sniderman, Paul M and Sean M Theriault. 2004. "The structure of political argument and the logic of issue framing." *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change* pp. 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 et al. 2007. "The exaggerated effects of advertising on turnout: The dangers of self-reports." *Quarterly Journal of Political Science* 2(4):325–343.