Does Reason-Giving Affect Political Attitudes? *

Jack Blumenau *University College London*

What are the effects of reason-giving on political attitudes? Both political philosophers and political scientists have speculated that defending proposals with reasons may change voters' preferences. However, while prominent models of attitude formation predict that the explicit justification of one's political views may result in attitudes that are more ideologically consistent, less polarized, and more stable, empirical work has not assessed the connection between reason-giving and attitudes. Implementing a survey experiment in which some respondents provide reasons before stating their opinions on six issues in UK politics, I find that reason-giving has very limited effects on the ideological constraint, temporal stability, or polarization of the public's political attitudes. These findings have important implications for our understanding of deliberative conceptions of democracy – in which reason-giving is a central component – as well as for our understanding of the quality of voters' political opinions.

Keywords: Public opinion; political attitudes; deliberation; reason-giving; survey experiments

---

9675 words

*__This version__: July 13, 2023. With thanks to Lucy Barnes, Peter Dinesen, Timothy Hicks, J. Scott Matthews and participants in seminars at University College London, Royal Holloway, Durham University, the University of Manchester, and the 2022 annual conference of the Elections, Public Opinion and Parties specialist group. This research was supported by funding from the British Academy (SRG2021-210655) and the Leverhulme Trust (RF-2021-327).

1 Introduction

Political discussion requires that, beyond stating the positions they hold, people articulate reasons for their policy preferences. Reason-giving is central to contemporary accounts of liberal theory (Habermas, 2015; Rawls, 1997; Chambers, 2010) and democratic deliberation (Thompson, 2008; Mutz, 2008; Gutmann and Thompson, 2009), which suggest that political legitimacy stems from the justifications that citizens and politicians give for their political choices. In addition to these intrinsic benefits, scholars have also speculated that the process of justifying one's political attitudes may also affect the content of those attitudes. As Cohen (2007, 228) suggests, "the practice of defending proposals with reasons may change my preferences." However, while recent work demonstrates that voters are able and willing to provide substantive reasons in defense of their preferences (Colombo, 2018, 2021), very little existing research evaluates whether and how reason-giving affects the attitudes that voters express. This is a significant omission given that reason-giving is considered to be the "first and most important characteristic" of deliberative democracy (Gutmann and Thompson, 2009, 24).

To understand how reason-giving might affect attitudes, I consider two models of attitude formation. On the one hand, models from political behaviour – which I refer to as *reasons-as-causes* models – hold that voters' reasons play a causal role in the construction of their attitudes (e.g. Azjen, 1980; Chong and Druckman, 2007; Nelson, Oxley and Clawson, 1997; Zaller, 1992; Zaller and Feldman, 1992) and suggest mechanisms through which reason-giving might lead to attitude change. In particular, by encouraging voters to engage in a greater degree of introspection about their preferences, these perspectives suggest that reason-giving may increase the *ideological constraint* and *stability* of voters' attitudes, and reduce *polarization* in the positions that voters adopt. Together, then, these models suggest that the heightened introspection induced by reason-giving could strengthen core properties of voters' democratic attitudes. On the other hand, models from social and political psychology – which I term *reasons-as-rationalization* models – suggest that reasons are used only to justify attitudes after they have been adopted (e.g. Lodge and Taber, 2013; Mercier and Sperber, 2018; Haidt, 2001). If rea-

sons are used to rationalise (rather than cause) political beliefs, this weakens the mechanisms through which articulating reasons might lead to attitude change. As a consequence, these models predict much more limited effects of reason-giving on political attitudes.

Theoretical disagreement about the predicted effects of reason-giving suggests a productive opportunity for empirical work. I describe an experimental design and associated analysis strategy aimed at providing evidence on the effects of reason-giving on political attitudes. In this design, survey respondents report their preferences on a set of political issues. While half of respondents provide *only* their policy preferences, the other half first provides the *reasons* that underpin their policy positions via an open-ended response, a treatment designed to increase cognitive effort and introspection. Having provided these justifications, the treatment-group answers the same set of policy questions as the control-group. I evaluate the effects of reason-giving by measuring differences between treatment and control with respect to three properties which represent different components of attitude "quality" (Price and Neijens, 1997) and which are predicted by *reasons-as-causes* models to be affected by reason-giving: constraint (the correlation between respondents positions on different issues, e.g. Achen, 1975; Zaller and Feldman, 1992; Ansolabehere, Rodden and Snyder, 2008); stability (the correlation of responses across survey waves, e.g. Achen, 1975; Zaller and Feldman, 1992; Freeder, Lenz and Turney, 2019); and polarization (the disagreement across respondents on each issue, e.g. Abramowitz and Saunders, 2008; Fiorina and Abrams, 2008).

Fielding this (pre-registered) experiment in a new two-wave panel survey of more than 5,000 UK citizens, I find that there are very limited effects of reason-giving on political attitudes. Despite some heterogeneity at the level of individual issues, reason-giving has precisely-estimated null average effects on both the polarization and stability of voters' attitudes, a finding that replicates across two survey samples. For constraint, attitudes are marginally more highly correlated across issues for the reason-giving respondents than for the control group in one sample of respondents, but this finding does not replicate in a second sample of respondents. For all three outcomes, I show that the null average effects do not mask significant heterogeneity between different subgroups of voters and that

these nulls are unlikely to be driven by a weak treatment. Taken together, the results demonstrate that providing justifications for one's political attitudes has no appreciable effects on the stability, constraint, or polarization of public opinion.

These findings have important implications for our understanding of both deliberative democracy and the quality of voters' political opinions. First, the experiment helps to open up the "black box of deliberation" (Mutz, 2008, 531) by examining the effects of one important feature of deliberation – reason-giving – on a particular set of political outcomes. Deliberative democrats have invoked a wide variety of requirements for successful deliberation, including civility, face-to-face exchange, and equality of participation, in addition to reason-giving. While a number of studies demonstrate the broader effects of deliberation on voters' attitudes (e.g. Gastil and Dillard, 1999; Sturgis, Roberts and Allum, 2005; Fishkin et al., 2020; Farrar et al., 2010; List et al., 2013; Minozzi et al., 2023), the deliberative experiences that form the basis of these studies include highly compound treatments, where reason-giving is bundled together with many other features of deliberation. By contrast, I demonstrate that one particular component of deliberative practice – the articulation of reasons – has essentially no effect on the attitudes that voters express. Focusing on individual deliberative actions is consistent with calls to further develop our understanding of whether, how, and under which conditions deliberation works. As Mutz (2008, 531) suggests, by "sorting out the importance of its contents in relation to various consequences, empirical research could greatly enhance the capacity of deliberative theory to contribute to democratic society".

Second, the results also speak to the relative efficacy of public versus private deliberation. Advocates of deliberative conceptions of democracy often suggest that inter-personal deliberation is critical for developing civic-virtues such as open-mindedness, tolerance, and attitude change (Dryzek, 2002; Rawls, 1997; Gutmann and Thompson, 2009; Cohen, 2005). For others, by contrast, deliberation between people is one mechanism by which voters might be encouraged to engage in "internal-reflective" deliberation and it is in this introspective reasoning that the true value of deliberation resides (Goodin, 2000; Goodin and Niemeyer, 2003). If the types of behavioural and attitudinal change

that have been associated with inter-personal deliberative experiences could be achieved by people deliberating alone, then the benefits of deliberation might be more easily scaled to more people. However, my results suggest that – at least with respect to the stability, constraint, and polarization of political attitudes – solitary reason-giving does not have effects equivalent to public deliberation. Consistent with recent work which demonstrates the importance of public over individual deliberation (Minozzi et al., 2023), the findings here suggest that private, internal, and individual reason-giving is not alone sufficient to produce the benefits ascribed to broader deliberative practices.

Finally, decades of survey research has painted a pessimistic picture about the capacity of voters to hold politicians to account on the basis of well-formed issue preferences (see Achen and Bartels, 2017, 30-36 for a review). Deficiencies in the political thinking of ordinary citizens – as evidenced, in part, by unstable, incoherent and polarized public attitudes – are taken to imply that conventional defenses of democratic government are "at odds with demonstrable, centrally important facts of political life" (Achen and Bartels, 2017, 306). Some hope that the quality of voters' expressed attitudes would increase if only voters could be induced to "think harder" about their political opinions, and evidence from other domains does suggest that greater mental effort and introspection can lead to more stable and coherent attitudes (e.g. Petty and Briñol, 2011). The results here, however, demonstrate that increased cognitive effort does not appear to result in such salutary effects for political attitudes. Despite the intrinsic value of reason-giving as a mechanism for legitimizing political decision-making, it is unlikely that more introspective and reason-based processing of political issues will alone act as a panacea to the problem of low-quality democratic attitudes.

2 Reason-Giving and Political Attitudes

Reason-Giving and Normative Democratic Theory

The idea that voters and elites are morally obligated to provide reasons for their political beliefs and choices is central to liberal democratic theory (Rawls, 1997; Chambers, 2010; Habermas, 2015). In

these accounts, reason-giving is typically seen as a mechanism through which political legitimacy is achieved and the ideals of mutual respect and the equality of persons are manifested. Rawls (1997, 771), for instance, argues that the moral obligation to engage in public reasoning about politics arises because the exercise of political power over others is only justified when "we sincerely believe that the reasons we would offer for our political actions…are sufficient." This proposition – that the public expression of reasons is intimately connected to democratic legitimacy – is most ardently articulated by theorists endorsing deliberative conceptions of democracy. The centrality of the "reason-giving requirement" (Gutmann and Thompson, 2009, 24) in deliberative democracy stems from the idea that presenting and responding to public reasons is the "primary conceptual criterion for [political] legitimacy" (Thompson, 2008, 504).

In addition to these intrinsic virtues, deliberation is also thought to "change minds and transform opinions" (Chambers, 2003, 318). Given the important role it plays in deliberation, reason-giving is seen as a central mechanism through which such effects might operate. For instance, Cohen (2007, 228) argues that "the practice of defending proposals with reasons may change my preferences". Similarly, Gutmann and Thompson (2009, 20) argue that when they deliberate, citizens engage in a process in which "the reasons given, and the reasons responded to, have the capacity to change minds." Moreover, changes in attitudes induced by deliberation are thought to be unambiguously positive. Voters who engage in deliberation are thought to be, *inter alia*, more willing to compromise, more tolerant of political difference, more consistent in their political attitudes, and more understanding of their own and others' political opinions (Mutz, 2008; Bortolotti, 2009; Cohen, 2005, 2007; Minozzi et al., 2023). These behavioural expectations therefore also reinforce the potential normative importance of reason-giving as they suggest that if inducing voters to engage in a reasoning process solicits political attitudes of higher "quality", then deliberative processes that prioritise reason-giving might serve to improve the prospects for democratic accountability.

Why might reason-giving affect the political attitudes that voters express? For many scholars, the mechanism by which deliberation is thought to affect preferences is through the public and social

exchange of views between people with different opinions (e.g. Dryzek, 2002; Cohen, 2005). However, other scholars have focused on the "internal-reflective" nature of deliberation in which the weighing of reasons "ultimately must take place within the head of each individual" (Goodin, 2000, 81). From this perspective, deliberation between individuals is valuable, in part, because it can induce voters to deliberate *internally*, giving more careful thought to their political attitudes. Reason-giving might therefore affect preferences via the introspection it engenders in voters (Goodin and Niemeyer, 2003). For instance, focusing on the idea that reason-giving might encourage introspection, Cohen (2005, 349) suggests that "the discovery that I can offer no persuasive reasons on behalf of a proposal of mine may transform the preferences that motivate the proposal". Similarly, Bortolotti (2009, 642) argues that reason-giving might shift attitudes by allowing people to develop a stronger understanding of their own beliefs and prompting them to think about the ideological coherence of their different attitudes. However, these accounts do not (and are not designed to) clearly articulate how introspection might affect attitudes, nor do they provide specific predictions about which properties of political attitudes might be affected by reason-giving. In the next section I therefore draw on models of attitude formation from political behaviour (*reasons-as-causes* models) and political and social psychology (*reasons-as-rationalization* models) which generate contrasting expectations for the effects of reason-giving on particular features of political attitudes.

Expected Effects of Reason-Giving on Political Attitudes

Reasons-as-causes models assume that voters form attitudes by averaging over a set of reasons relevant to a given issue and that reported attitudes are determined by those reasons. Crucially, in this perspective, reasons play a *casual* role in opinion formation: if the set of reasons that a voter considers relevant to a given issue changes, then the voter's opinion on that issue may also change. In a prominent example of such an argument, Zaller (1992) suggests that voters have in their heads a distribution of potentially competing "considerations" from which they sample stochastically when prompted to express their political opinions on a given subject. Attitude reports do not therefore represent the

considered opinions of voters on particular issues, but rather reflect the outcome of a process in which voters average over those sampled considerations and make choices "in great haste – typically on the basis of the one or perhaps two considerations that happen to be at the 'top of the head' at the moment of response" (Zaller, 1992, 36). The idea that attitudes are causally determined by aggregating across reasons is shared by other accounts (e.g. Azjen, 1980; Nelson, Oxley and Clawson, 1997; Chong and Druckman, 2007), but it is the intuition that voters draw a *sample* of reasons each time they are required to produce a political opinion, and it is from this sample that they construct their attitudes, that drives many of the interesting theoretical predictions of the effects of reason-giving discussed below.

By contrast, *reasons-as-rationalization* models see political attitudes as deriving from fast and intuitive processing in which explicit reasoning plays a very limited role. Rooted in "dual-process" understandings of human cognition (Evans, 2008), these models suggest that the slow, deliberative, and conscious evaluation of reasons ("System-2" thinking) is likely to be rare, with most attitudes forming as a result of fast, automatic and unconscious processes ("System-1" thinking). Lodge and Taber (2013), for instance, argue that voters do not consider and evaluate political arguments and justifications in order to form preferences, but rather that voters' attitudes arise from the affect-driven processes that are characteristic of System-1 thinking. The idea that people will provide evaluations without engaging in a conscious, inferential reasoning process is also common in both social (Mercier and Sperber, 2018) and moral (Haidt, 2001) psychology. Even when voters have the time, motivation and opportunity to engage in deliberative reasoning, these perspectives suggest that the process of reasoning will itself be biased by the fast and unconscious judgments that are driven by the valence of initial affect towards a given issue. In these models, then, reasons are used by voters to *rationalise* their intuitively formed attitudes. As Mercier and Sperber (2018, 112) suggest, reasons do not "motivate or guide us in reaching conclusions" but rather "justify after the fact the conclusions we have reached." *Reasons-as-rationalization* models therefore differ sharply from *reasons-as-causes* models, as the causal path connecting reasons to attitudes runs in reverse: people produce reasons to support the attitudes

they intuitively adopt, rather than constructing their attitudes from the reasons they hold.

What do these models predict for the effects of reason-giving on political attitudes? I focus on three properties of voters' attitudes which have been interpreted as relating to the "quality" of public opinion (Price and Neijens, 1997). First, *reason-as-causes* models suggest that reason-giving might have important effects on the *stability* of voters attitudes. If, following Zaller, voters form attitudes by sampling from a population of reasons, then the variance of voters' attitudes will be lower when the voter draws a larger sample of considerations.[1] As Zaller puts it, "responses formed by averaging over a larger number of considerations would be better indicators of the population of underlying considerations than responses based on just one or two considerations, and hence more reliable" (Zaller, 1992, 86). As a consequence, we should expect attitudinal instability – the degree to which voters' attitudes change over time – to be lower in contexts where they are induced to think about a wider range of considerations related to a given policy. Zaller argues that the key to increasing the number of considerations used in forming attitudes is increased engagement or "extra thought" (Zaller, 1992, 86) about a given issue. A similar argument can be found in the "elaboration likelihood model" of attitude change (e.g. Petty and Briñol, 2011), in which attitude strength, stability and coherence are seen as a function of the amount of thought that people devote to a given attitude object. As Petty and Briñol (2011) suggest, "the more a judgment is based on thinking about the merits of an issue, the more it tends to persist over time." Therefore, if reason-giving provokes voters to "slow down and reexamine his or her line of thought" Mansbridge (2007, 262), then we should expect justification-providing voters to express more stable attitudes than voters who are not asked to provide reasons for their attitudes.

[1]Consider a voter i forming an attitude towards policy p and time t ($V_{i,p,t}$) as function of a set of J "considerations", $v_j^{p,t}$, that the voter holds about that policy:

$$V_{i,p,t} = \frac{1}{J}\sum_{j=1}^{J} v_j^{p,t}$$

If the $v_j^{p,t}$ considerations used to evaluate policy p are sampled from a broader distribution with variance $\sigma_{i,p}^2$ then $V_{i,p,t}$ has variance $Var(V_{i,p,t}) = \frac{\sigma_p^2}{J}$, implying that variability in expressed policy preferences is a decreasing function of the number of considerations sampled (i.e. J).

Second, reason-giving might also increase the correlation between attitudes on different political issues – a quantity typically referred to as attitude *constraint* (Converse, 1964). One key mechanism driving this prediction is again that the sampling variation of attitudes will be related to the effort exerted in searching for reasons. The correlation between voters expressed attitudes on different issues will be biased towards zero when the variance of those attitudes is high. Therefore, if reason-giving induces voters to consider a larger number of reasons when constructing attitudes, their expressed attitudes will be less variable, and the correlation of their attitudes across issues will increase.

A second, more substantive, mechanism linking reason-giving to constraint is that providing justifications might also make voters aware of conceptual links across different issues, thus inducing them to express more correlated attitudes. Explicitly stating the justifications for issue A might make them more present in the minds of respondents when thinking about issue B. For instance, if a voter believes that "the poor don't have enough to get by" is an important justification for their support for a higher tax rate on high-income individuals, then the articulation of that belief might encourage them to recognise the potential validity of the same justification when considering a subsequent question about unemployment benefits. Similarly, if a voter believes that "individuals should be free to make their own choices" is a valid defense of their views on free speech, articulating that justification might make it a more prominent feature in determining their attitudes towards transgender rights. If voters who think about the reasons for their beliefs are more likely to make connections between issues that have common underpinnings, they may therefore be more likely to express correlated views on those topics.

Finally, reason-giving might also affect the *polarization* of voters' attitudes. Again, this might result from different mechanisms. First, reason-giving could – à la Zaller – increase the number of sampled considerations and reduce the variance of expressed attitudes which would, in expectation, result in less polarized attitudes across voters on a given issue. This moderating effect occurs purely as a result of the reduced variability in attitudes that comes from averaging over a larger set of considerations. Second, engaging in reason-giving might also induce voters to consider the arguments

on the other side of the issue more carefully, thus encouraging them to take a more moderate position on the issue. For instance, Goodin (2000, 98) suggests that "[t]hrough the exercise of a suitably informed imagination, each of us might be able to conduct a wide-ranging debate within our own heads among all the contending perspectives." This idea is central to many "perspective-taking" accounts of political moderation, which suggest that understanding the experiences and perspectives of political opponents can durably reduce political polarization (Kalla and Broockman, 2022, 2020; Broockman and Kalla, 2016).

These expectations rely on models of attitude formation that view reasons as causes, but what does the *reasons-as-rationalization* model predict for the effects of reason-giving on attitudes? For the most part, this perspective suggests that reason-giving should have little or no effect on expressed attitudes. If attitudes are determined by affective, intuitive and unconscious responses to external stimuli, and reasons are used to post-hoc justify spontaneously generated feelings, then this considerably weakens the mechanism through which thinking about and articulating those reasons can lead to attitude change. Crucially, for these accounts, any cognitive reasoning process about an object will be biased by the initial affective response to that object which reduces the probability that introspection about reasons will shift attitudes. As a result, we should expect introspective reason-giving to have very limited effects on attitudes.[2] One exception to this pattern relates to attitudinal polarization, where *reason-as-rationalization* perspectives suggest that reason-giving might actually *increase* polarization. If people reason in a biased manner, then the initial affective reaction they have in response to any given object might be further reinforced by the accumulation of reasons that align with that response. As Mercier and Sperber (2018, 247) suggest, when a person engages in such motivated reason-giving, it will "increase her confidence and lead her to extreme positions."

In summary, the arguments in this section demonstrate the significant theoretical ambiguity over the expected effects of reason-giving on political attitudes. Reasons-as-causes models suggest posi-

[2] Rational choice models also predict that solitary reason-giving will not affect political attitudes because it does not reveal new information to agents. However, while these models provide similar predictions as the *reasons-as-rationalization* model, they do not offer a psychological basis for such predictions (Dietrich and List, 2016).

tive effects of reason-giving on the ideological constraint and temporal stability of political attitudes, and negative effects of reason-giving on polarization. By contrast, reasons-as-rationalization models predict no effect of reason-giving on stability or ideological constraint, and positive effects on polarization.

Empirical Evidence on the Effects of Reason-Giving

Existing evidence from social and cognitive psychology suggests that engaging in processes of reasoning can affect the attitudes people endorse (e.g. Tesser, 1978). In particular, introspecting about reasons appears to affect the decisions that people take and the satisfaction they subsequently feel from those decisions (Wilson and Schooler, 1991; Wilson et al., 1993; Dijksterhuis, 2004; Simonson, 1989; Hsee, 1999). The broad conclusion of this literature is that "people who reason more act differently from those who reason less or not at all" (Mercier and Sperber, 2018, 253). However, while these studies offer *prima facie* evidence that exerting additional mental effort in evaluating an object can affect attitudes, they do not directly address reason-giving as a specific mechanism for attitude change. Moreover, many of outcomes in these papers relate to consumers' choices over products, which limits the degree to which they are informative about properties of political attitudes.

In political science, existing evidence suggests that voters participating in inter-personal deliberative forums develop attitudes that are more ideologically constrained (Sturgis, Roberts and Allum, 2005; Gastil and Dillard, 1999) and less polarised (Fishkin et al., 2020), are more likely to change their minds on an issue (Minozzi et al., 2023), and also have preferences that come closer to demonstrating properties of single-peakedness (Farrar et al., 2010; List et al., 2013) than voters who did not participate in those forums. However, the deliberative settings that underpin these studies represent highly compound treatments, as – in addition to reason-giving – participants also receive a great deal of policy-relevant information, engage in group-based discussion, cast votes for preferred outcomes, and so on. Therefore, while these initiatives are helpful for determining whether deliberation *as a whole* affects attitudes, they are not informative about the effects of *individual elements* of deliberation,

such as reason-giving. If reason-giving is thought to affect attitudes in particular ways, the appropriate test is one which compares the views of those who engage in reason-giving to those who do not. As Mutz (2008, 530) suggests, to understand the mechanisms that drive the effects of deliberation, we need to "identify which characteristics of deliberative practice produce which kinds of desirable outcomes", a sentiment shared by many other scholars (e.g. Gastil and Dillard, 1999, 21; Thompson, 2008, 500-501).

The study that comes closest to evaluating the effects of reason-giving on attitudes is by Zaller and Feldman (1992)[3] who randomly assigned some survey respondents to answer a "stop-and-think" question which required them to report some relevant considerations before providing their views on a given issue. Consistent with the discussion above, Zaller and Feldman expected respondents in the stop-and-think condition to report attitudes that were more stable across survey waves and more highly correlated across issues. However, stopping-and-thinking increased ideological constraint only for respondents with high levels of political sophistication, while attitude stability was (insignificantly) *lower* in the stop-and-think condition than in the control condition (Zaller and Feldman, 1992, 605).

Despite some similarities with the design below, the experiment reported in Zaller and Feldman (1992) represents an incomplete test of the effects of reason-giving. First, the treatment administered by Zaller and Feldman (1992) was a thought-listing exercise,[4] which is conceptually distinct both from the treatment described below and from reason-giving as understood in the literature on deliberation. Second, the experiment was fielded as a part of the 1987 ANES pilot study to a very small sample of respondents (only 450 respondents in the first wave, and 357 in the second), making the null results somewhat difficult to interpret. Third, their analysis focused on only three issues, which limits the generalisability of the findings. Finally, the response options available to respondents differed between the treatment and control groups, a decision that reintroduces the possibility of selection bias.

[3] Also reported in Zaller (1992, 85-89).

[4] "Before telling me how you think about this, could you tell me what kinds of things come to mind when you think about [POLICY]?"

As a result of these issues, Zaller (1992, 91) concluded that the predictions of his model that relate to the effects of reasoning on constraint and stability "cannot be said to have been adequately tested." Further research into the effects of reason-giving on attitudes therefore seems warranted.

3 Experimental Design

In this section, I describe the design of a two-wave online panel survey which was fielded to UK respondents by Opinium in early 2022. All analyses described below were pre-registered with the Evidence in Governance and Politics (EGAP) registry [REDACTED FOR PEER REVIEW].[5]

Sample and Randomization

The first survey wave – fielded in January 2022 – consisted of 3010 respondents, who were selected using nationally representative quotas for gender, age, vote in the 2019 UK General election and political attention. In the first survey wave, respondents were randomly assigned into two groups with equal probability. Respondents in each group were asked to report their positions on four issues (sampled at random from a set of 6 issues, described below) in current UK politics. Respondents in the control group were *only* asked to provide their preferred policy option on each issue. Respondents in the treatment group were asked, before giving their policy preferences, to provide the reasons for their positions on each issue (prompt described below). After providing their reasons, treatment-group respondents then answered the same set of policy questions as the control group. I refer to results from the first sample of respondents in the first wave of the survey as "Sample One, Wave One" results.

2545 respondents from the first wave were successfully recontacted in the second survey wave, fielded in May and June 2022. These respondents were asked to provide their preferences (and, if in the treatment group, reasons) for the same set of political issues that they considered in wave one. The treatment assignment persisted across the two waves of the survey such that reason-giving respondents in wave one also provided reasons for their positions in wave two. This allows me to

[5] An anonymised version of the pre-analysis plan is appended to this document.

assess the extent to which repeated treatment exposure affects expressed attitudes. I refer to results from this set of respondents as "Sample One, Wave Two" results.

In addition, the second wave also included 1438 new respondents who did not appear in the first wave. These newly added respondents in wave two were also randomized into treatment and control groups with equal probability and followed the same survey as other wave two respondents (with the four issues sampled at random). This allow me to replicate two of the analyses (for constraint and polarization) on a fresh sample. I refer to results from this second sample of respondents as "Sample Two" results.

Policy Areas

The six policies included in the experiment included a mix of high- and low-salience issues, including four broadly related to the economic "left-right" dimension of UK politics ("Unemployment Support", "Higher Rate of Tax", "Minimum Wage" and "Zero hours contracts") and two related to the social "liberal-conservative" dimension ("Transgender Rights" and "Offensive Speech"). These issues also span a range of "easy" (symbolic and easily-communicable) issues and "hard" (technical and complex) issues (Carmines and Stimson, 1980), attitudes on which are thought to be structured by different types of cognitive processes.[6] Several of the policies were drawn from those used in Hanretty, Lauderdale and Vivyan (2020), while others were written to cover more recently topical issues in UK politics. Each respondent answered questions relating to four out of the six issues. Each issue was paired with a thematically similar issue (discussed below) and sampling was conducted at the issue-pair level, such that for each respondent two issue-pairs were sampled and respondents provided responses to all four issues.

Although the experiment results in a reasonable number of observations for each policy/treatment-group combination, the design is only sufficiently powered (see appendix section B) to detect relatively

[6]Voters' attitudes on "easy" issues are thought to be governed by their "gut responses", while preferences on "hard" issues are the result of "a sophisticated decision calculus" (Carmines and Stimson, 1980, 78).

large treatment effects at the level of individual issues.[7] An alternative design would have been to select a smaller number of issues and gather a larger number of responses for each of them. However, that approach would be subject to generalizability concerns, as any inferences would be limited to the specific issues included. Instead, I use a larger number of policy areas, but focus on the average effect of the treatment across issues. Using a large set of policy issues maximizes the external validity of the experimental results, while targeting the average effect of the treatment effect maximizes the power of the design (Blumenau and Lauderdale, 2022).

Survey Prompts

Figure 1 provides an example of the open-ended reason-giving prompt displayed to respondents in the treatment group for the "Higher Rate of Tax" issue. After a short introduction, respondents were asked to provide the reasons that supported their view on whether the government should increase or decrease the rate of income tax for high-income individuals. This prompt was designed to reflect how reason-giving is conceived in the theoretical literature. First, consistent with Mansbridge (2007, 261), who argues that " 'reason-giving' can include any statement that sincerely answers the 'why' question", the prompt instructs voters to provide the reasons that they see as supporting their own position on the issue. Second, by asking respondents to "think very carefully" about their own reasons, it provides a plausible inducement for respondents to engage in the type of "internal-reflective process" that many scholars believe is a key mechanism linking deliberation to attitude change (Goodin, 2000, 95; see also Bortolotti, 2009; Cohen, 2005; Goodin and Niemeyer, 2003). Finally, the prompt emphasises that respondents should "explain as many reasons as possible for your view", a phrase which directly attempts to manipulate the number of considerations that respondents draw into their minds at the point of attitude formation, something that is central to many of the predictions of the *reasons-as-causes* model (Zaller, 1992; Zaller and Feldman, 1992). It is worth stressing that treatment group

[7] For instance, the 3010 "Sample One, Wave One" respondents would result in approximately 1003 observations for each policy-by-treatment-group combination, and just 848 and 479 for the "Sample One, Wave Two" and "Sample Two" respondents, respectively.

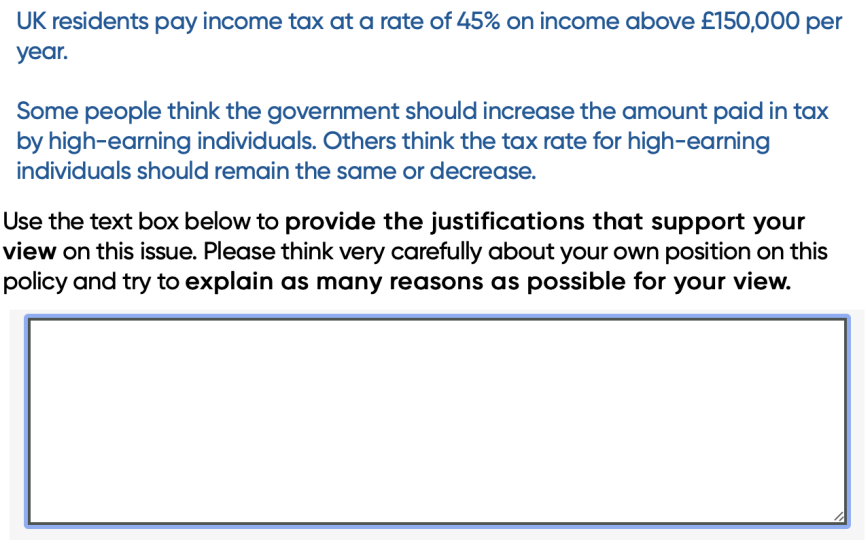


Figure 1: Reason-giving prompt

respondents did not receive any additional information relative to the control group. Rather, the treatment aims to provoke the type of introspection that the *reasons-as-causes* model predicts will be consequential for attitude change.

After providing justifications, the treatment group were asked to select the position closest to their own from five logically ordered alternatives (plus a "Don't know" response option). Figure 2 provides an example for the "Higher Rate of Tax" issue. In this case, respondents could select a taxation rate for yearly incomes above £150,000, with options ranging from ten percentage points below to fifteen percentage points above the current status quo (45%).

Control-group respondents, by contrast, saw *only* the issue-position prompt in figure 2. The full text of both prompts for each of the six issues included in the experiment is given in appendix A.

4 Measuring Constraint, Stability, and Polarization

To assess the effects of reason-giving on attitudes, I analyse the correlation between responses on different issue items (*constraint*), the correlation on the same issue items across survey waves (*stability*), and the dispersion of responses across respondents on each item (*polarization*). I discuss the measure-

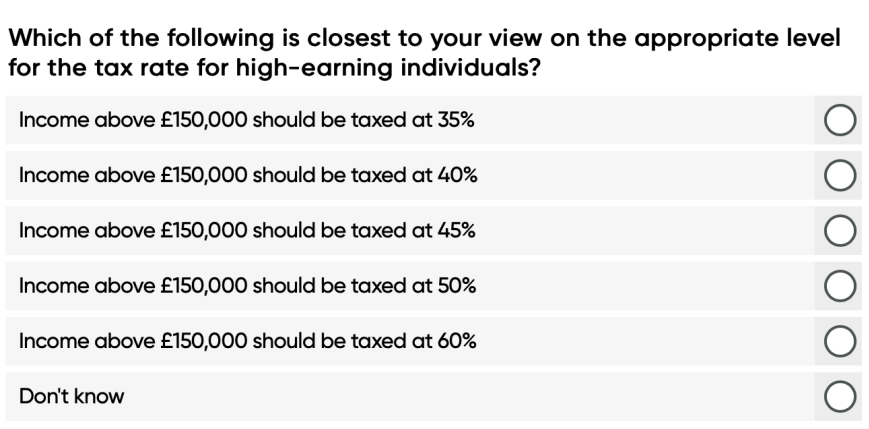


Figure 2: Issue position prompt

ment of these quantities, all of which have been widely employed in the existing literature, below. As declared in the pre-registration plan, I 1) conduct all analyses using survey weights; 2) recode the policy item variables such that higher scores indicate more left-wing or more socially-liberal positions; and 3) remove "Don't know" responses for any of the policy questions.[8]

Constraint

To investigate the effects of reason-giving on ideological constraint, I measure the degree to which correlations between issue stances are higher in the reason-giving treatment group than in the control group. In particular, I calculate the weighted polychoric correlation between each pair of policy items for each group, where, because all policy items are recoded to indicate more left-wing responses, higher correlations indicate a greater degree of ideological consistency across items. The differences in these correlations for each issue-pair (e.g., $\rho^{D=1}_{\text{HighTax,MinWage}} - \rho^{D=0}_{\text{HighTax,MinWage}}$) reflect the extent to which the reason-giving treatment induces more highly correlated attitudes *on a given pair of issues* relative to the control condition. However, as noted above, the design is well-powered to detect only large treatment effects at the individual issue level, and so for each group I also calculate the

[8]Averaging across issues in the first wave of the survey, 14% of responses were "Don't know" responses, with a minimum of 7% for the unemployment support issue and a maximum of 22% for the offensive speech issue. Treatment group respondents were 1.25 percentage points more likely to provide a "Don't know" response than control group respondents, on average, though this difference is insignificant ($t = 1.64$, standard errors clustered at the respondent level).

average correlation across the 15 issue-pairs. The main inferential quantity of interest is therefore the difference in these average correlations between treatment and control groups (i.e. $\overline{\rho}^{D=1}_{\text{Constraint}} - \overline{\rho}^{D=0}_{\text{Constraint}}$). When this difference is positive, it suggest that reason-giving respondents report attitudes that are more consistently left- or right-wing across issues compared to control-group respondents.

In addition, the theoretical discussion revealed that we should expect the effects of reason-giving to differ across different pairs of the six issues included in the experiment. One mechanism through which the effects of reason-giving might operate is by making respondents aware of common justifications that apply across different, but related, political issues. For instance, common reasons might support a respondent's views on both the "minimum wage" and "zero hours contracts" issues, but it is less likely that common reasons would apply to the "higher rate of tax" and "transgender rights" issues. Evidence for this mechanism therefore requires categorising the pairs of issues that plausibly have common substantive underpinnings. Before fielding the experiment, I selected 3 pairs of issues that I expected to "hang together" in terms of their underlying ideological stance. These pairings were as follows:

1. Increase Unemployment Support/Increase Higher Rate of Tax
2. Increase Minimum Wage/Restrict Zero Hours Contracts
3. Expand Transgender rights/Limit Offensive Speech

These pairings reflect an expectation that attitudes on issues of this sort *could* be underpinned by common reasons. If the effects of reason-giving run primarily through an increased appreciation of arguments that are common across policies, we should expect effects to be stronger for these selected pairs of policies than for other issue pairs. I preregistered this expectation and highlight estimates from these selected issue-pairs in the results below.

Stability

To measure the stability of voters' attitudes, I calculate weighted polychoric correlations of the six policy items between survey waves for both treatment and control groups. These correlations capture

the degree to which respondents' answers in the first wave of the survey persisted in the second wave of the survey. The differences in the correlations for each issue (e.g. $\rho^{D=1}_{\text{HighTax}} - \rho^{D=0}_{\text{HighTax}}$) therefore reflect the extent to which respondents in the treatment group ($D = 1$) have more or less stable attitudes for a given issue than respondents in the control group ($D = 0$). As with the constraint measure, the main quantity of interest is the difference in the *average* (i.e. across issue) correlations ($\bar{\rho}^{D=1}_{\text{Stability}} - \bar{\rho}^{D=0}_{\text{Stability}}$) between treatment and control groups.

Polarization

To measure the polarization of issue-based preferences, I calculate the weighted mean absolute error (MAE) of the responses to each policy item in the treatment (e.g. $MAE^{D=1}_{\text{HighTax}}$) and control groups (e.g. $MAE^{D=0}_{\text{HighTax}}$).[9] The MAE is the average of the absolute differences between each survey response and the sample mean, meaning that higher values of the MAE indicate that responses to a given policy item are more polarized.[10] As with the other measures, in addition to reporting issue-level treatment effects, the main inferential focus is on the average difference in MAE across issues between treatment and control groups ($\overline{MAE}^{D=1} - \overline{MAE}^{D=0}$). Positive values for this difference indicate that the average polarization of attitudes is higher in the treatment group and negative values indicate higher average polarization in the control group.

For all quantities of interest, I evaluate sampling uncertainty via a non-parametric bootstrap. I resample 500 times from the original survey data with replacement, blocking on individual respondents, and I construct the quantities above for each iteration. I summarise the results of this procedure using 95% confidence intervals for all quantities.

[9] For respondents $i \in 1, ...N$ in groups $d \in 0, 1$, on issues $k \in 1, ..., K$, the MAE is given by:

$$MAE^{D=d}_k = \frac{1}{\sum w_i} \sum_{i=1}^{N_{D=d}} w_i |\mu^{D=d}_k - X^k_i|$$

where X^k_i is the response on issue k by respondent i, μ_k is the mean survey response on issue k and w_i is a survey weight.

[10] In appendix section F I demonstrate that the substantive conclusions are unaffected by using alternative measurement strategies for polarization.

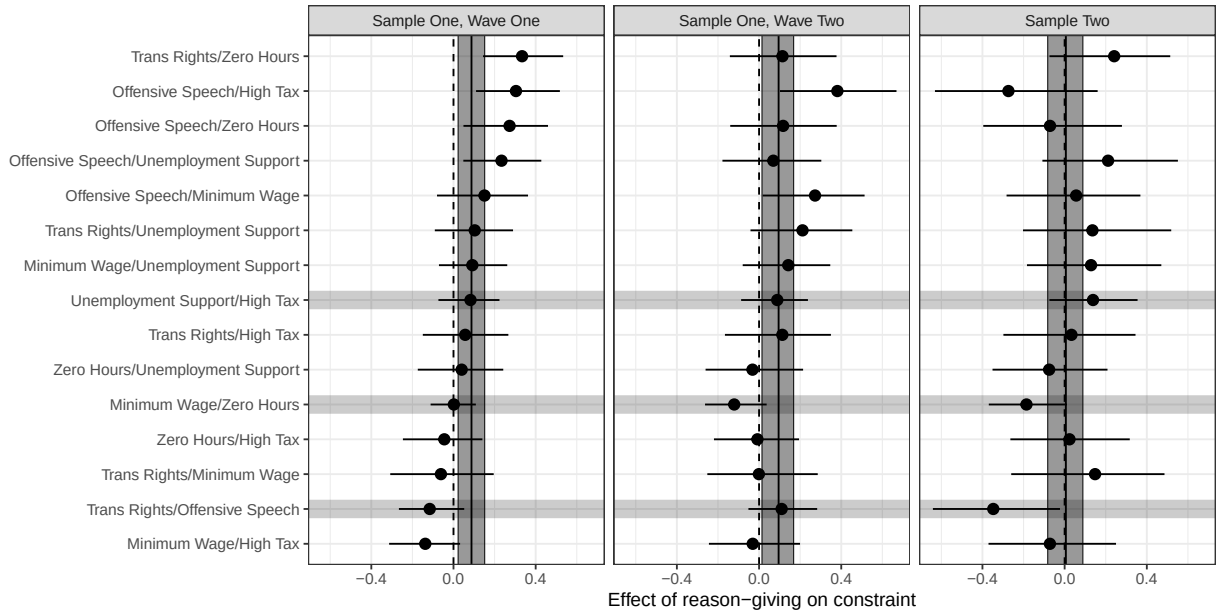


Figure 3: Effects of Reason-Giving on Constraint

5 Results

Constraint

Figure 3 depicts the estimated treatment effects for all 15 pairwise correlations between the 6 issues included in the experiment. The left and centre panels of the figure show the effects for the first sample of respondents, with correlations measured in the first and second waves of the survey, respectively. The right-hand panel shows the effects for the second sample of respondents. Points further to the right indicate that reason-giving respondents had attitudes that were more highly correlated on a given pair of issues than control-group respondents. Points further to the left indicate that the control group responses were more highly correlated. Vertical lines represent the average treatment effects across issues for each sample/survey wave.

The figure indicates that reason-giving results in a small average increase in the correlation of attitudes across issues for the first sample of respondents: the average correlation across issues for respondents in the treatment group was 0.086 [0.013, 0.149] points higher than for those in the control

group. The average effect of reason-giving is also roughly the same magnitude after the treatment is repeated in the second wave of the survey, where the estimated difference between treatment and control respondents is 0.100 [0.019, 0.175]. However, this effect does not replicate in the second sample of respondents, where the estimated treatment effect is 0.005 [-0.067, 0.090]. Taken together, these results – which average across the effects on different issue pairs – provide only weak support for the idea that reason-giving induces people to provide more ideologically consistent responses.

In addition, the figure also reveals significant heterogeneity in the effects of reason-giving across the issue-pairs included in the experiment. For instance, for the "Sample One, Wave One" results, the estimated treatment effect for the Trans Rights/Zero Hours issue-pair was 0.334 [0.144, 0.534], which implies that treatment-group responses on these issues were marked by substantially higher correlations than control group responses. By contrast, on the Minimum Wage/High Tax issue-pair, the estimated treatment effect was -0.137 [-0.313, 0.032], implying that those proving reasons for their preferences reported attitudes that were somewhat *less* correlated than those in the control group.

Notably, the positive effects of justification on constraint do not appear to be driven by the pairs of issues which I expected, *a priori*, to be more responsive to reason-giving. The gray horizontal bars in figure 3 indicate the issue pairs that were selected as being thematically related in the pre-analysis plan. If the effects of reason-giving run primarily through an increased appreciation of arguments that are common across policies, then we should expect effects to be stronger for policies that are thematically related than for those that address very different underlying rationales. However, as the figure reveals, the effects of reason-giving are actually *smaller* for these issue pairs than the average treatment effect across all issue pairs. Across these three issues, the average effect of reason-giving was indistinguishable from zero for the first sample of respondents in both wave one (-0.010 [-0.083, 0.064]) and wave two (0.026 [-0.069, 0.119]), and negative (though insignificant) for the second sample of respondents (-0.132 [-0.278, 0.032]). Somewhat surprisingly, the largest effects of reason-giving appear for issue pairs that include both the first and second dimensions of British politics. For instance, when voters give reasons for their policy views, attitudes on the two social issues (transgender rights

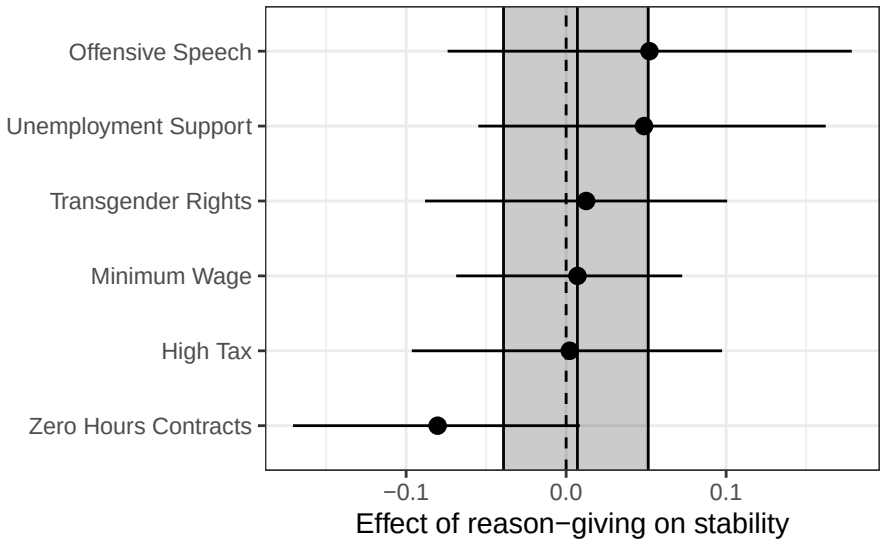


Figure 4: Effects of Reason-Giving on Stability

and offensive speech) become more correlated with attitudes on a number of economic issues, such as zero hours contracts, unemployment support and the minimum wage. Again, however, these patterns are mostly present for the first sample of respondents but do not replicate in the second sample, making it hard to put a lot of weight on these inferences.

Stability

Figure 4 presents the estimated effects of reason-giving on the stability of public attitudes. I again present estimates for each issue included in the experiment, and the main quantity of interest – the average effect of the treatment across all issues – is depicted with vertical lines and confidence bands. As stability is only measurable for the set of respondents who appear in both waves of the survey, I present only one set of estimates for this outcome variable.

As with the constraint analysis, despite some heterogeneity at the issue-level, the average effect of the reason-giving treatment on the stability of expressed attitudes is close to zero (0.007 [-0.039, 0.051]). For none of the individual issues is the treatment effect significant and positive, and in one case – the Zero Hours Contracts issue – reason-giving appears to decrease attitude stability relative to the control group. This evidence therefore again fails to conform to the prediction of the *reasons-*

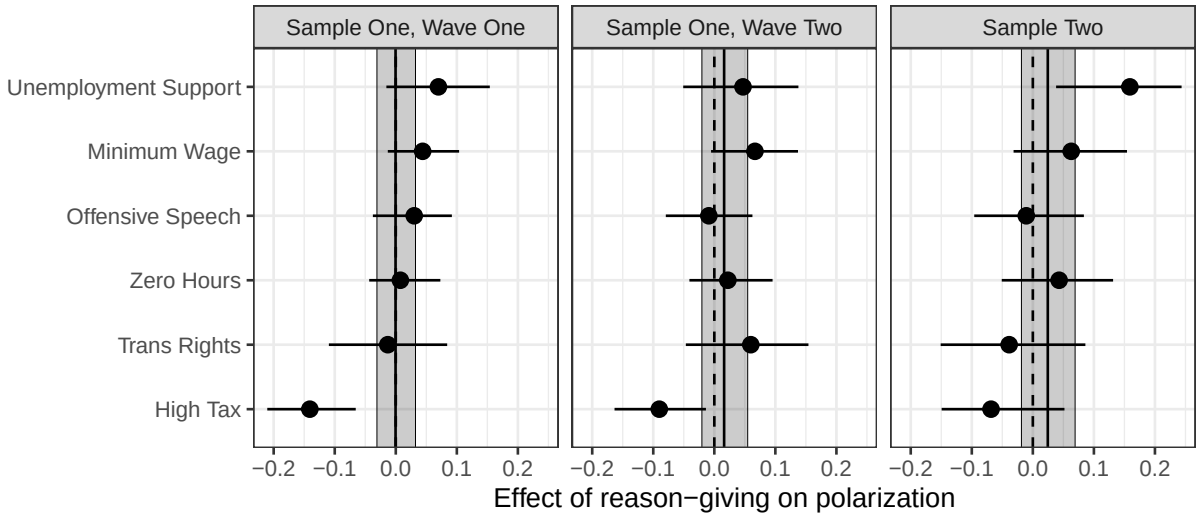


Figure 5: Effects of Reason-Giving on Polarization

as-causes model that reason-giving will lead to greater attitudinal stability. That does not appear to be the case here, as people engaged in reason-giving have attitudes that demonstrate as much temporal variation as those who do not provide reasons for their attitudes.

Polarization

Finally, figure 5 shows the estimated difference in the mean absolute error between treatment-group and control-group respondents on each of the six issues included in the experiment. Again, vertical lines and error bars indicate the average effects across issues, and I present estimates for the different samples of respondents and the different waves of the survey.

By now, the story is familiar: there is a reasonably large amount of treatment heterogeneity across issues but the average effect of the treatment is very close to zero. For example, reason-giving appears to modestly increase attitude polarization on the unemployment support issue, but modestly decreases the polarization of attitudes on the appropriate rate of tax for high-income individuals. More importantly, the average effect of reason-giving on attitude polarization is very close to zero. For respondents in the first sample, the average treatment effect is indistinguishable from zero in both wave one (0.000 [-0.031, 0.033]) and wave two (0.016 [-0.021, 0.049]) of the survey. The same is true

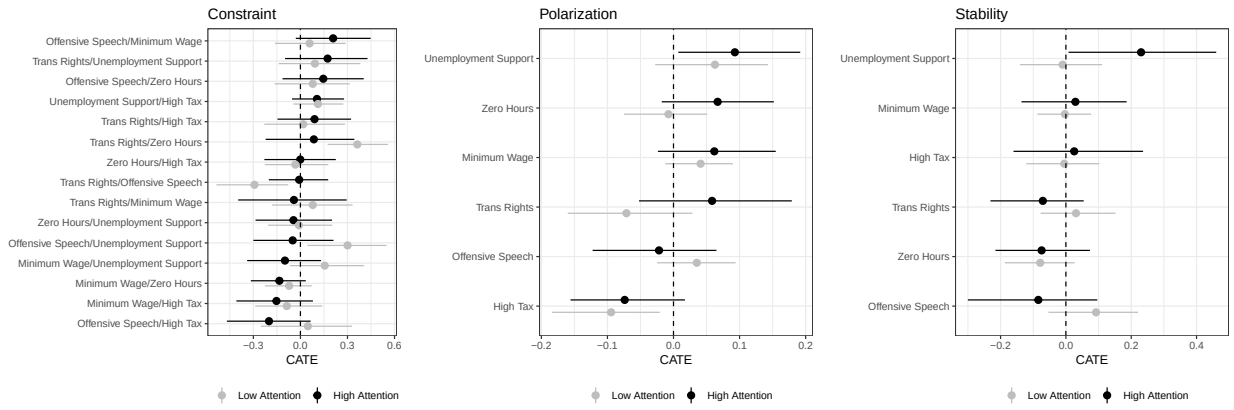


Figure 6: Conditional Issue-Level Treatment Effects by Political Attention

for the second sample of respondents where the estimated treatment effect is 0.024 [-0.020, 0.065]. Together, these results again fail to support the idea that reason-giving might have systematic effects on political attitudes.

Heterogeneous Treatment Effects

Do these null average effects mask heterogeneity at the respondent level? One might expect, for instance, that the effects of reason-giving would to be more pronounced for voters who typically exert little effort thinking about politics (e.g. Zaller, 1992, 86-88). For such voters, engaging in reason-giving could have strong effects because it is for these voters that greater introspection might most expand the set of considerations brought to mind. By contrast, for voters who typically pay more attention to politics, reason-giving could have less pronounced effects because such voters are likely to already consult a broad variety of considerations when forming their opinions. To test this expectation, figure 6 presents issue-level treatment effects, conditional on respondents' self-reported level of political attention.[11]

There is little evidence that the effects of reason-giving vary systematically by political attention. Although for some issues and issue-pairs there are small differences between the treatment effects for

[11]Political attention is measured on an 11-point scale, ranging from "0 - Pay no attention" to "10 - Pay a great deal of attention" (plus a "Don't know" option). As pre-registered, I divide respondents into two groups according to whether they are above or below the median response on this variable. In order to maximise power, for the constraint and polarization outcomes, I pool together responses from the first and second samples for these subset analyses.

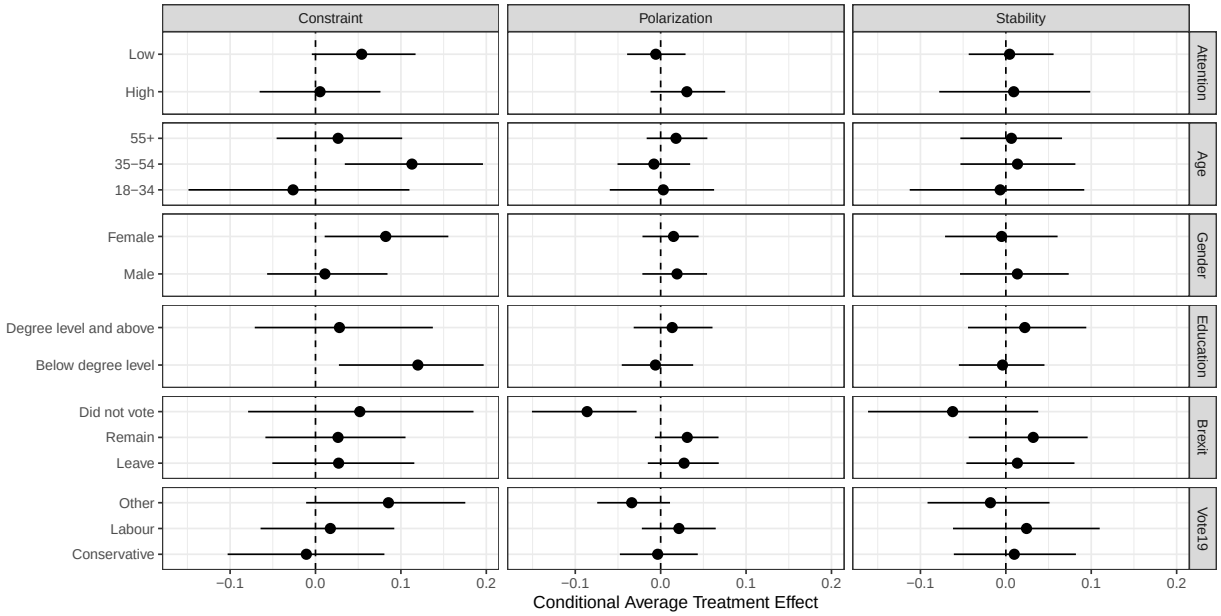


Figure 7: Conditional Average Treatment Effects by Respondent Characteristics

high- and low-attention respondents, in general there is a high degree of correlation across issues and it is not the case that low-attention respondents are systematically more responsive to the treatment than other respondents.

In analyses that were not pre-registered, figure 7 shows the *average* (i.e. across issues) effect of the reason-giving treatment on each outcome for a number of different groupings of respondents, determined by age, gender, education, political attention, and past vote in the 2016 Brexit referendum and the 2019 general election. The figure reveals that there is, in general, relatively little treatment-effect heterogeneity. For the stability outcome, the results are especially uniform, with null effects of reason-giving across all groups of respondents. Similarly, for the polarization outcome, providing justifications for one's attitudes has effects that are indistinguishable from zero for all groups except those who did not vote in the 2016 referendum. For this group, I estimate a small negative effect of the reason-giving treatment. For the constraint outcome, there is also limited evidence of treatment-effect heterogeneity. Lower-education respondents are somewhat more affected by the treatment, as are women and those aged between 35 and 54, but these differences are small in magnitude.

Taken together, these results suggest that the average effects reported above do not mask highly

differential responses to the treatment by different groups of respondents. Both on average and within demographic groups, reason-giving has very limited effects on the stability, constraint and polarization of attitudes that voters express.

6 Threats to Inference

One potential objection to the analyses above is that people may not have engaged sufficiently with the reason-giving treatment. If the treatment did not provoke respondents to think more deeply about their attitudes, then the null effects above might be attributable to the experimental design rather than reflecting properties of the attitude formation process. I provide two pieces of evidence that inconsistent with this "weak treatment" interpretation.

First, there is clear evidence that reason-giving respondents spend more time thinking about a given issue before providing their responses than do control-group respondents. Figure A2 in appendix section C shows the amount of time in seconds that respondents spent on the introductory screen for each issue, which they viewed before providing their issue preferences. For control group respondents, who only saw a short introduction to the issue, the median time spent contemplating the issue before providing their preferences was 6.0 [5.0, 7.0] seconds. By contrast, reason-giving respondents – who saw the same introduction to the issue as the control group but then also provided justifications – spent 69.0 [67.1, 70.9] seconds contemplating the issue before stating their preferences. That is, the typical treatment-group respondent spent over a minute longer – a ten-fold increase – thinking about the issue at hand before providing their policy preferences than did the typical control-group respondent.

Second, the content of the reasons provided by respondents in the treatment group suggests a high degree of engagement with the underlying policy issues. The median length of responses to the open-ended reason-giving prompt was between 15 and 22 words, depending on the policy issue, which provides reassuring face validity that respondents were engaging with the reason-giving task.

In addition, in appendix H, I provide evidence that the reasons people provided via the open-ended survey response are substantively related to the issues under consideration and that supporters and opponents of different policy positions use predictably different words in justifying their personal stances (figure A13). This again suggests that people were following the instructions in the prompt and actively considering the reasons that lie behind their political beliefs. Taken together, then, these findings suggest that it is unlikely that the reason-giving treatment was so weak that it failed to compel respondents to canvas their minds for salient considerations.

An additional threat to inference is the possibility of bias stemming from differential item and unit non-response across the treatment and control groups. The reason-giving treatment requires (by design) greater cognitive effort than the control condition, which could cause less motivated respondents in the treatment group to refuse to answer some questions or drop out of some survey waves altogether. If such non-response is associated with the any of the outcomes – stability, polarization or constraint – it would cause the estimated treatment effects above to be biased. For all three dependent variables, the likely direction of non-response bias would make it more likely to find effects of reason-giving on attitudes. Given that the treatment requires respondents to engage in a effortful political-reasoning task, it is plausible that non-response in the treatment-group will be higher among less politically-sophisticated voters and voters who are less interested in politics. These voters are also likely to have attitudes that are less ideologically structured, more polarized, and less stable, on average (e.g. Converse, 1964; Zaller, 1992; Zaller and Feldman, 1992; Ansolabehere, Rodden and Snyder, 2008; Freeder, Lenz and Turney, 2019). As a consequence, estimated treatment effects are likely to be upwardly biased, as respondents who remain in the treatment-group sample are those for whom we would expect higher levels of constraint and stability and lower levels of polarization. Given the likely direction of the bias, it is all the more striking that the results here suggest such limited effects of reason-giving. In addition, in appendix D, I replicate the main analyses in the paper using inverse-probability-of-attrition weights (IPAWs) to adjust for differential item and unit non-response (Gerber and Green, 2012). I show that the substantive findings reported here are not

sensitive to the incorporation of such weights.

In appendix section E, I also demonstrate that the null results are very unlikely to be attributable to ceiling or floor effects. While there is variation in the constraint, polarization and stability outcomes across issues for the control group, it is not the case that the response distributions for these quantities is either so high or so low that shifts to those distributions are impossible. These results suggest that the limited effects of reason-giving are therefore not an artifact of the outcome questions used in the survey.

Finally, readers might wonder whether reason-giving has effects on properties of attitudes other than constraint, stability or polarization. Most obviously, one plausible hypothesis is that reason-giving respondents might provide responses that are systematically more or less liberal or conservative, or further to the left or the right, on a given issue. In appendix G, I show that there is no consistent evidence of such effects. Although some differences appear on individual issues – reason-giving respondents report attitudes that are further to the right on the High Tax and Unemployment Support issues – the magnitude of these differences is very small, and the average effect of reason-giving across all issues is indistinguishable from zero.

7 Conclusion

The exchange of arguments in favour of and in opposition to different policy alternatives is a defining feature of political discussion and deliberation. The core contribution of this paper is to show that such reason-giving does not, in isolation, have the salutary effects on political attitudes hoped for by some proponents of deliberative democracy. Stability, constraint, and polarization are all important aspects of voter preferences because of the role they play in strengthening democratic accountability and facilitating political agreement. Although reason-giving is only one feature of deliberative exchange, it is often seen as a central mechanism through which deliberation can result in such properties becoming manifest. Consistent with calls to investigate "important, specifiable, and falsifiable

parts of deliberative democratic theory" (Mutz, 2008, 521), evaluating the effects of reason-giving in isolation is therefore an important endeavour. The findings here – that explicit justification of one's preferences does not affect the expression of those preferences – do not undermine the claim that deliberation, *in toto*, might have beneficial effects on democratic attitudes, but they nevertheless provide important evidence about one of the core behavioural assumptions that underpins the deliberative turn in normative theory.

The findings here also imply that whatever weaknesses exist in the political attitudes of the public, "fixing individual reasoning is not the solution" (Mercier and Landemore, 2012, 254). That is, simply inducing voters to devote more cognitive effort to the reasons that underpin their attitudes is insufficient for improving the quality of those attitudes. However, while the treatment employed here – solitary, introspective reason-giving, divorced from broader discursive context – aimed to solicit the type of "internal-reflective" reasoning that is seen as critical by many scholars (e.g. Goodin, 2000; Goodin and Niemeyer, 2003; Petty and Briñol, 2011), it might miss potentially important effects stemming from *public* political reasoning. Rawls (1997, 786), for instance, argues, "Public justification is not simply valid reasoning, but argument addressed to others." Similarly, Mercier and Sperber (2018) argue that reasoning evolved as a response to problems encountered in social interaction and that the most valuable aspects of reason-giving are social. Future work should therefore investigate whether different *types* of reason-giving have effects on political attitudes, and under which conditions. For example, a more powerful reason-giving task might be one in which voters are asked to persuade and respond to the reasons of other people, rather than just engaging in introspection. Similarly, future work might examine the effects of reason-giving on other types of attitudinal outcome. The exchange of reasons between voters of different political opinions, for example, might help to decrease hostility and increase understanding across lines of political disagreement.

One strength of the design here is that it facilitated an assessment of treatment-effect heterogeneity across a variety of different issues, and the results revealed that the effects of reason-giving are not entirely uniform across issues. Given this heterogeneity, a reasonable concern is that the re-

sults reported here may not generalise to samples from other countries, or on other sets of political issues. External validity concerns are greatest, however, when there is strong reason to believe that treatment effects will vary with political context and it is not clear that we should expect a significant degree of heterogeneity in the effects of reason-giving on attitudes. The process of reason-giving is unlikely to be very different across countries, and it is hard to come up with a compelling theoretical rationale for why we might expect such treatments to have very different effects on UK voters versus voters in other places. Ultimately, however, the degree to which these results travel to other contexts is an empirical question, and future work should consider replicating analyses such as those reported here in other settings.

7 References

Abramowitz, Alan I and Kyle L Saunders. 2008. "Is polarization a myth?" *The Journal of Politics* 70(2):542–555.

Achen, Christopher H. 1975. "Mass political attitudes and the survey response." *American Political Science Review* 69(4):1218–1231.

Achen, Christopher H and Larry M Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government*. Vol. 4 Princeton University Press.

Ansolabehere, Stephen, Jonathan Rodden and James M Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review* 102(2):215–232.

Azjen, Icek. 1980. "Understanding attitudes and predicting social behavior." *Englewood cliffs* .

Blumenau, Jack and Benjamin Lauderdale. 2022. "The Variable Persuasiveness of Political Rhetoric." *American Journal of Political Science* .

Bortolotti, Lisa. 2009. "The epistemic benefits of reason giving." *Theory & Psychology* 19(5):624–645.

Broockman, David and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Carmines, Edward G and James A Stimson. 1980. "The two faces of issue voting." *American Political Science Review* 74(1):78–91.

Chambers, Simone. 2003. "Deliberative democratic theory." *Annual review of political science* 6(1):307–326.

Chambers, Simone. 2010. "Theories of political justification." *Philosophy Compass* 5(11):893–903.

Chong, Dennis and James N Druckman. 2007. "Framing theory." *Annual review of political science* 10(1):103–126.

Cohen, Joshua. 2005. Deliberation and democratic legitimacy. In *Debates in contemporary political philosophy*. Routledge pp. 352–370.

Cohen, Joshua. 2007. *Deliberation, Participation and Democracy*. Palgrave Macmillan chapter 10 - Deliberative Democracy.

Colombo, Céline. 2018. "Justifications and citizen competence in direct democracy: A multilevel analysis." *British Journal of Political Science* 48(3):787–806.

Colombo, Céline. 2021. "Principled or Pragmatic? Morality Politics in Direct Democracy." *British Journal of Political Science* 51(2):584–603.

Converse, Philip E. 1964. The nature of belief systems in mass publics. In *Ideology and Discontents*, ed. David Apter. Glencoe Free Press.

Dietrich, Franz and Christian List. 2016. "Reason-based choice and context-dependence: An explanatory framework." *Economics & Philosophy* 32(2):175–229.

Dijksterhuis, Ap. 2004. "Think different: the merits of unconscious thought in preference development and decision making." *Journal of personality and social psychology* 87(5):586.

Dryzek, John S. 2002. *Deliberative democracy and beyond: Liberals, critics, contestations*. Oxford University Press on Demand.

Evans, Jonathan St BT. 2008. "Dual-processing accounts of reasoning, judgment, and social cognition." *Annu. Rev. Psychol.* 59:255–278.

Farrar, Cynthia, James S Fishkin, Donald P Green, Christian List, Robert C Luskin and Elizabeth Levy Paluck. 2010. "Disaggregating deliberation's effects: An experiment within a deliberative poll." *British journal of political science* 40(2):333–347.

Fiorina, Morris P and Samuel J Abrams. 2008. "Political polarization in the American public." *Annu. Rev. Polit. Sci.* 11:563–588.

Fishkin, James, Alice Siu, Larry Diamond and Norman Bradburn. 2020. "Is deliberation an antidote to extreme partisan polarization? Reflections on America in One Room.".

Freeder, Sean, Gabriel S Lenz and Shad Turney. 2019. "The importance of knowing "what goes with what": Reinterpreting the evidence on policy attitude stability." *The Journal of Politics* 81(1):274–290.

Gastil, John and James P Dillard. 1999. "Increasing political sophistication through public deliberation." *Political communication* 16(1):3–23.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Goodin, Robert E. 2000. "Democratic deliberation within." *Philosophy & Public Affairs* 29(1):81–109.

Goodin, Robert E and Simon J Niemeyer. 2003. "When does deliberation begin? Internal reflection versus public discussion in deliberative democracy." *Political Studies* 51(4):627–649.

Gutmann, Amy and Dennis F Thompson. 2009. *Why deliberative democracy?* Princeton University Press.

Habermas, Jürgen. 2015. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108(4):814.

Hanretty, Chris, Benjamin Lauderdale and Nick Vivyan. 2020. "The Emergence of Stable Political Choices from Incomplete Political Preferences." *Working Paper* .

Hsee, Christopher K. 1999. "Value seeking and prediction-decision inconsistency: Why don't people take what they predict they'll like the most?" *Psychonomic Bulletin & Review* 6:555–561.

Kalla, Joshua L and David E Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments." *American Political Science Review* 114(2):410–425.

Kalla, Joshua L and David E Broockman. 2022. "Voter Outreach Campaigns Can Reduce Affective Polarization among Implementing Political Activists: Evidence from Inside Three Campaigns." *American Political Science Review* pp. 1–7.

List, Christian, Robert C Luskin, James S Fishkin and Iain McLean. 2013. "Deliberation, single-peakedness, and the possibility of meaningful democracy: evidence from deliberative polls." *The journal of politics* 75(1):80–95.

Lodge, Milton and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

Mansbridge, Jane. 2007. *Deliberation, Participation and Democracy*. Palgrave Macmillan chapter 12 - "Deliberative Democracy" or "Democratic Deliberation"?

Mercier, Hugo and Dan Sperber. 2018. The Enigma of Reason. In *The Enigma of Reason*. Penguin Random House.

Mercier, Hugo and Hélène Landemore. 2012. "Reasoning is for arguing: Understanding the successes and failures of deliberation." *Political psychology* 33(2):243–258.

Minozzi, William, Ryan Kennedy, Kevin M Esterling, Michael A Neblo and Ryan Jewell. 2023. "Testing the Benefits of Public Deliberation." *American Journal of Political Science* .

Monroe, Burt L, Michael P Colaresi and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4):372–403.

Mutz, Diana C. 2008. "Is deliberative democracy a falsifiable theory?" *Annu. Rev. Polit. Sci.* 11:521–538.

Nelson, Thomas E, Zoe M Oxley and Rosalee A Clawson. 1997. "Toward a psychology of framing effects." *Political behavior* 19(3):221–246.

Petty, Richard E and Pablo Briñol. 2011. "The elaboration likelihood model." *Handbook of theories of social psychology* 1:224–245.

Price, Vincent and Peter Neijens. 1997. "Opinion quality in public opinion research." *International Journal of Public Opinion Research* 9(4):336–360.

Rawls, John. 1997. "The idea of public reason revisited." *The University of Chicago Law Review* 64(3):765–807.

Simonson, Itamar. 1989. "Choice based on reasons: The case of attraction and compromise effects." *Journal of consumer research* 16(2):158–174.

Sturgis, Patrick, Caroline Roberts and Nick Allum. 2005. "A different take on the deliberative poll: Information, deliberation, and attitude constraint." *Public Opinion Quarterly* 69(1):30–65.

Tesser, Abraham. 1978. Self-generated attitude change. In *Advances in experimental social psychology*. Vol. 11 Elsevier pp. 289–338.

Thompson, Dennis F. 2008. "Deliberative democratic theory and empirical political science." *Annu. Rev. Polit. Sci.* 11:497–520.

Wilson, Timothy D, Douglas J Lisle, Jonathan W Schooler, Sara D Hodges, Kristen J Klaaren and Suzanne J LaFleur. 1993. "Introspecting about reasons can reduce post-choice satisfaction." *Personality and Social Psychology Bulletin* 19(3):331–339.

Wilson, Timothy D and Jonathan W Schooler. 1991. "Thinking too much: introspection can reduce the quality of preferences and decisions." *Journal of personality and social psychology* 60(2):181.

Zaller, John R. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Zaller, John and Stanley Feldman. 1992. "A simple theory of the survey response: Answering questions versus revealing preferences." *American journal of political science* pp. 579–616.

Appendix Table of Contents

Contents

A Survey Prompts

Box 1: Higher Rate of Tax

UK residents pay income tax at a rate of 45% on income above £150,000 per year.

Some people think the government should increase the amount paid in tax by high-earning individuals. Others think the tax rate for high-earning individuals should remain the same or decrease.

Treatment group only:

*Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view**.*

[TEXT BOX]

Which of the following is closest to your view on the appropriate level for the tax rate for high-earning individuals?

- *Income above £150,000 should be taxed at **35%***
- *Income above £150,000 should be taxed at **40%***
- *Income above £150,000 should be taxed at **45%***
- *Income above £150,000 should be taxed at **50%***
- *Income above £150,000 should be taxed at **60%***
- *Don't know*

Box 2: Unemployment Support

Some people think the government should provide unemployment benefits to people whenever they are out of work. Others think that unemployment benefits should be provided for limited periods or that the government should not provide such benefits at all.

Treatment group only:

Use the text box below to ***provide the justifications that support your view*** *on this issue. Please think very carefully about your own position on this policy and try to* ***explain as many reasons as possible for your view****.*

[TEXT BOX]

Which of the following is closest to your view on the appropriate level of support that the government should provide for UK citizens of working age who are not employed?

- ***People should be paid unemployment benefit whilst they are out of work.*** *This unemployment benefit should last as long as the person is unemployed.*
- ***People should be paid unemployment benefit whilst they are out of work.*** *This unemployment benefit should last* ***as long as the person is unemployed, and as long as they can show that they are actively seeking a job.***
- ***People should be paid unemployment benefit in their first few months out of work only.***
- ***People should not generally be paid unemployment benefit, except where they are unable to work because of a disability or injury they got whilst working.***
- ***There should be no unemployment benefit.*** *Individuals unable or unwilling to find work should be supported by family, friends, or charities.*
- *Don't know*

Box 3: Minimum Wage

Some people think that the government should increase the minimum wage in the UK. Others think that the government should maintain, or even reduce, the minimum wage.

Treatment group only:

*Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view**.*

[TEXT BOX]

Which of the following is closest to your view on the appropriate level for the minimum wage?

- *The government should **remove the minimum wage entirely** and let businesses decide how much to pay workers.*
- *The government should **keep the minimum wage at the current level** (£8.91 per hour).*
- *The government should **increase the minimum wage by a small amount** (£9.50 per hour).*
- *The government should **increase the minimum wage by a larger amount** (£11 per hour).*
- *The government should **increase the minimum wage by a substantial amount** (£15 per hour).*
- *Don't know*

Box 4: Zero Hours Contracts

Some people think the government should take action to reduce or ban zero hours contracts (contracts with no guarantee of hours or income). Others think zero hours contracts should remain available as an option for employers.

Treatment group only:

*Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view**.*

[TEXT BOX]

Which of the following is closest to your view on on zero hours contracts (contracts with no guarantee of hours or income)?

- *Zero hours contracts **should be permitted** under whatever terms employers and employees agree to.*
- *Zero hours contracts **should be permitted, but employers should commit to employment hours at least one day in advance**, and pay wages when they cancel with less notice.*
- *Zero hours contracts **should be permitted, but employers should commit to employment hours at least one week in advance**, and pay wages when they cancel with less notice.*
- ***Workers on zero hours contracts should be subject to a higher minimum wage than normal contracts.***
- ***Zero hours contracts should be illegal.***
- *Don't know*

Box 5: Transgender Rights

Transgender people who wish to change their legal gender on official documents (e.g. birth certificate, passport, etc) have to apply for a Gender Recognition Certificate. This requires someone to have a diagnosis of gender dysphoria from a doctor, provide evidence that they have lived in their new gender for at least two years, and make a declaration that they intend to live in their new gender for the rest of their lives.

Some people think that the government should reduce the amount of documentation required for transgender people to change their gender on official documents. Others think that the government should increase the amount of documentation or not allow the gender on official documents to change at all.

Treatment group only:

*Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view**.*

[TEXT BOX]

Which of the following is closest to your view on the requirements for transgender people who wish to change their gender on legal documents?

- ***Transgender people should be able to change their gender on legal documents without providing any evidence at all.***
- *The government should **reduce the amount of evidence required** for transgender people to change their gender on legal documents.*
- ***The current requirements** for transgender people to provide evidence to change their gender on legal documents **are about right.***
- *The government should **increase the amount of evidence required** for transgender people to change their gender on legal documents.*
- ***Transgender people should not be allowed to change their gender on legal documents under any circumstances.***
- *Don't know*

Box 6: Offensive Speech

Some people think that the government should stop people from saying things that offend other people. Others think that the government should not ban offensive speech.

Treatment group only:

*Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view**.*

[TEXT BOX]

Which of the following is closest to your view on offensive/hate speech?

- *Government **should not stop people from saying offensive things**, no matter who is affected.*
- *Government should stop people from saying things that offend people of different **races**.*
- *Government should stop people from saying things that offend people of different **races or religions**.*
- *Government should stop people from saying things that offend people of different **races, religions, or sexual orientations**.*
- *Government should stop people from saying things that offend people of different **races, religions, sexual orientations, or political beliefs**.*
- *Don't know*

B Power Analyses

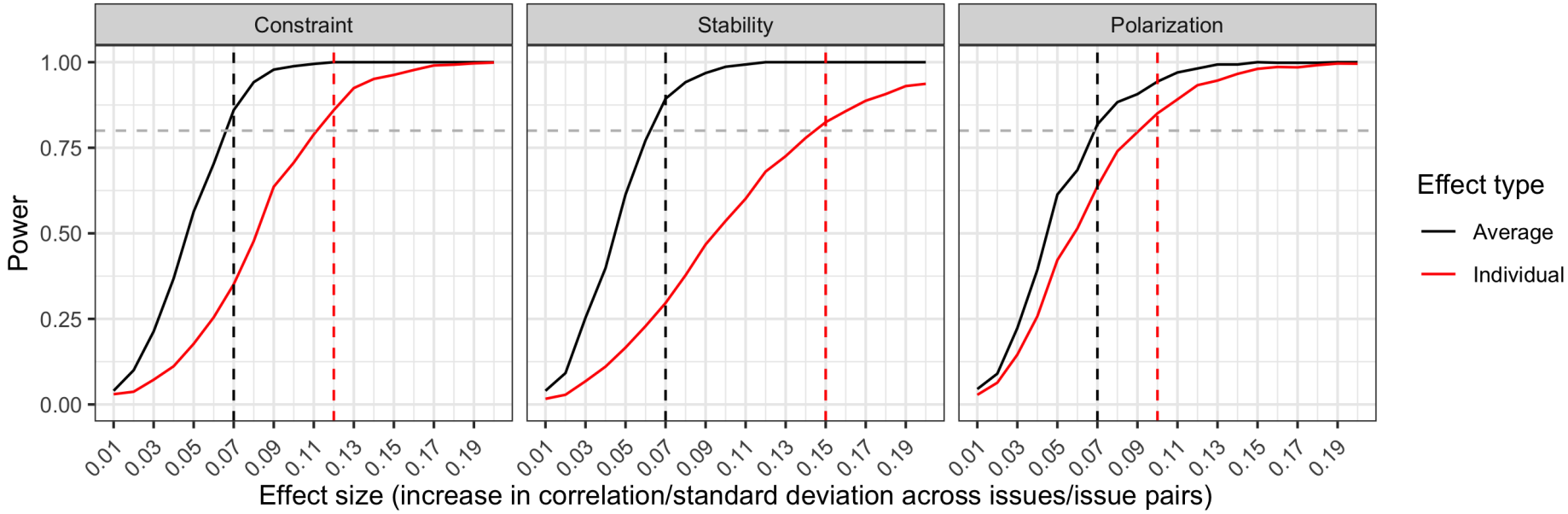
Figure A1 shows the results of a power analysis for the quantities of interest described in section 4 of the paper. To construct the power analysis, I simulated the data collection process for a fixed sample size ($N = 3000$), for four policy responses per respondent, and for different hypothetical treatment effects. For the stability analysis, I also assumed an attrition rate of 30% across survey waves (uncorrelated with the treatment).

Establishing a reasonable expectation for treatment effect magnitudes is difficult in this application because previous studies have not evaluated the effects of survey format on the correlation between policy items, on the stability of responses on items over time, or on the polarization of voter opinions. For the two correlation-based measures (stability and coherence), I used reasonably conservative hypothetical treatment effects, ranging from zero to an increase in the average correlation of 0.2. For the polarization measure, the effect size is measured in the difference in standard deviations of the response variable for the treatment and control groups.

The black lines in the figure depict the power for the average treatment effects described section 4 of the paper. The red lines in the figure represent the power for detecting treatment effects for *individual* policies (for the stability and polarization outcomes) and for policy pairs (for the constraint outcome). The minimum detectable effects (MDE) for a sample size of 3000 and a power of 0.8 are presented as vertical lines in each panel.

Figure A1 clearly illustrates that the design is only sufficiently powered to detect reasonably large effects for individual policies or policy pairs. The MDE for individual policy effects is 0.15 for the stability outcome and 0.1 for the polarization outcome. The MDE for individual policy-pair effects is 0.12 for the constraint outcome. By contrast, the MDEs for the average treatment effects are considerably smaller, at 0.07 for constraint, polarization and stability.

Figure A1: Power analysis



C Question Duration by Treatment Group

Figure A2 shows the amount of time in seconds that respondents spent on the introductory screen for each issue, which they viewed before providing their issue preferences. The typical treatment-group respondent spent over a minute longer – a ten-fold increase – thinking about the issue at hand before providing their policy preferences than did the typical control-group respondent.

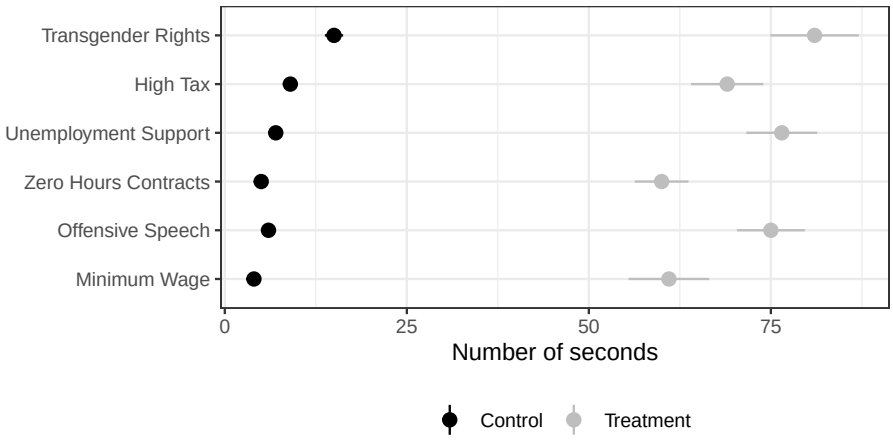


Figure A2: Introductory screen duration per issue

D Item and Unit Non-Response

As described in the main body of the paper, differential item and unit non-response between treatment and control groups could bias the estimates of the effects of reason-giving for all three dependent variables. There is evidence of differential item and unit non-response for the treatment and control groups in the data here. Of the 3383 respondents who began the first wave of the survey, 99% of control group respondents finished the survey compared to only 90% of treatment group respondents. Similarly, of the 1606 control respondents who completed the first wave of the survey, 77% also completed wave two, compared to just 68% of the 1404 treatment group respondents. If this non-response was also correlated with the constraint, polarization or stability of respondents' attitudes, then it is plausible that the estimates presented in the paper are subject to bias.

As argued in section 5 of the paper, bias of this form is overwhelmingly likely to lead to *over*-estimates the effects of reason-giving and is therefore (given the null results) unlikely to threaten the inferences drawn in the paper. However, it is nevertheless worth trying to establish the degree to which the estimates presented here are sensitive to these differential response patterns.

To do so, in this section I report robustness checks for each of the main analyses in the paper in which I estimate inverse-probability-of-attrition weights (IPAWs) to adjust for differential item and unit non-response. IPAWs measure the inverse of the probability of a given observation being observed in a given analysis, on the basis of observable covariates. IPAWs require estimating the relationship between attrition and the available covariates, constructing a probability of being observed for each unit, and then taking the reciprocal of that probability to form a weight (Gerber and Green, 2012, Chapter 7). The intuition behind this approach is that survey respondents with characteristics that are similar to the missing observations will be up-weighted in the analyses which will therefore mitigate the bias caused by attrition.

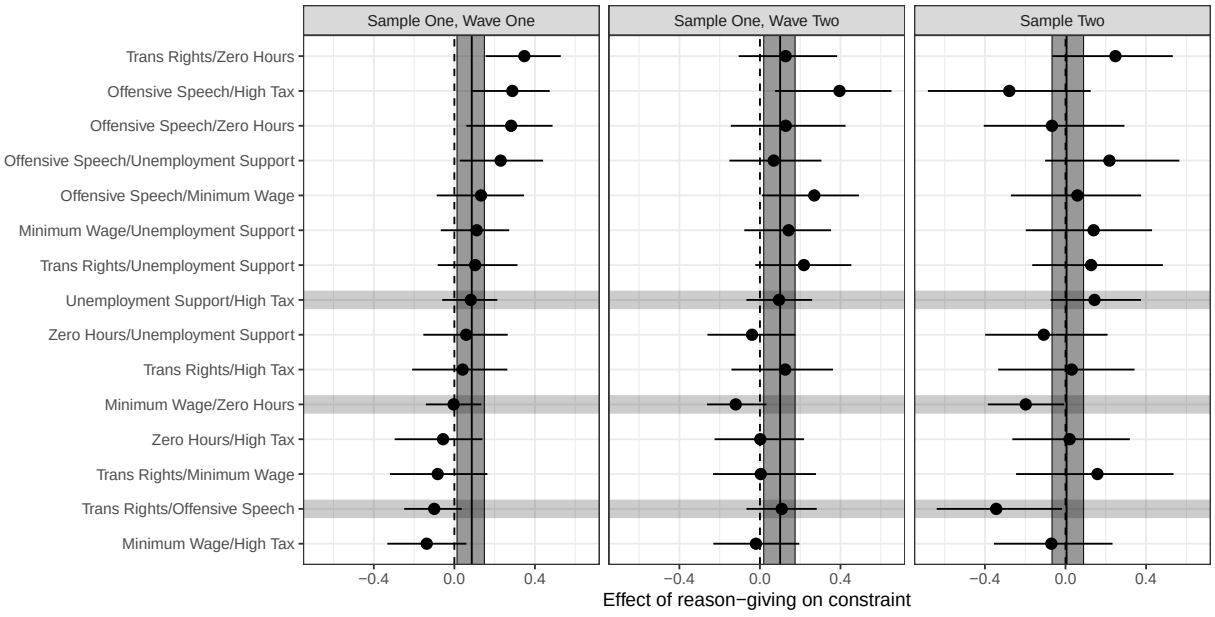


Figure A3: Effects of Reason-Giving on Ideological Constraint (Attrition Weighted)

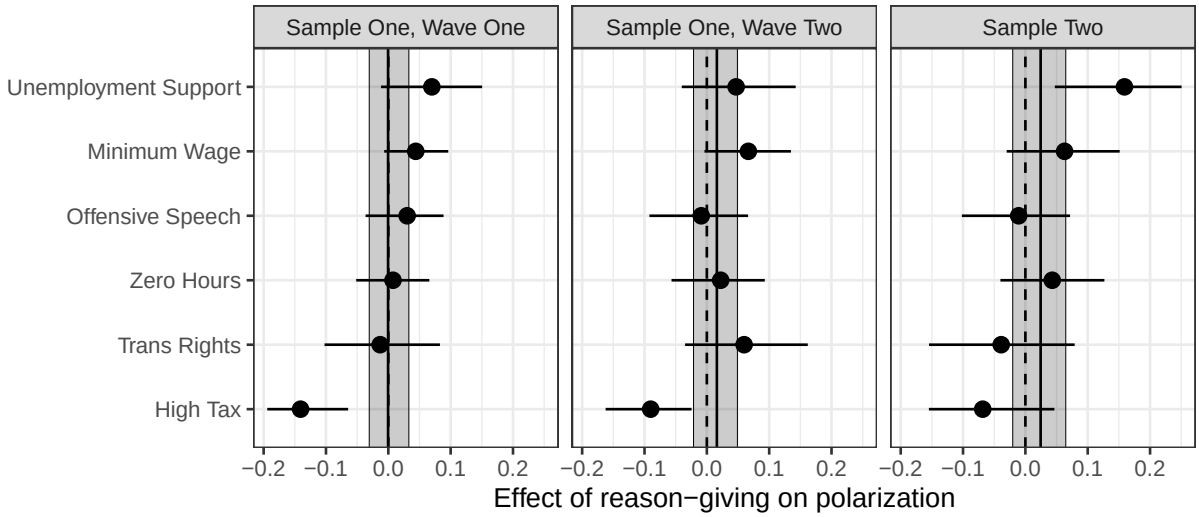


Figure A4: Effects of Reason-Giving on Attitude Polarization (Attrition Weighted)

I estimate IPAWs using logistic regression applied both to the responses within each wave (for the constraint and polarization outcomes) and across waves (for the stability outcome). For the within-wave weights, I estimate a logistic regression where the dependent variable is equal to one if a respondent completed the survey wave, and zero otherwise. I model this outcome as a function of age, gender, political attention, employment, education, vote in the 2019 general election, as well as interactions between each of those variables and the treatment indicator. For the across-wave weights, I estimate a logistic regression where the dependent variable is equal to one when a respondent from wave one also appeared in wave two, and zero otherwise. I use the same variables to model the relationship between being observed in both waves and respondent characteristics.

I use these probabilities to construct IPAWs, which I incorporate into the analysis (alongside the survey weights) and replicate the findings presented in the paper in figures A3, A4, and A5. As the results make clear, accounting for non-response does not have any substantive effect on the results. The effects of reason-giving on both polarization and stability of respondents' attitudes is zero, and there is a very small positive effect of reason giving on attitude constraint in the first sample, but not the second sample, of respondents.

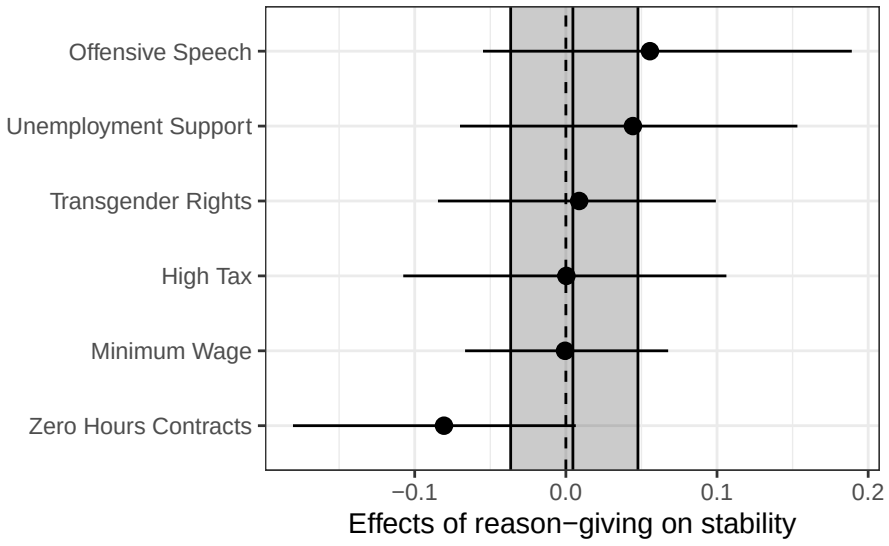


Figure A5: Effects of Reason-Giving on Attitude Stability (Attrition Weighted)

E Ceiling and Floor Effects

One potential concern is that the results reported in the paper might be attributable to ceiling or floor effects. If levels of constraint and stability are near their maximum for control group respondents, or levels of polarization are near their minimum, then my ability to detect changes in these response distributions would be limited. In this section, I therefore report the levels of the three main quantities of interest for both the treatment and control group.

Constraint: Figure A6 depicts the treatment- and control-group correlations between issue positions on each of the 15 pairs of issues included in the experiment. Positive values on the x-axis indicate that left (right) responses on one issue tend to be accompanied by left (right) responses on the other issue in a pair, while negative correlations indicate that left (right) responses on one issue tend to go together with right (left) responses on the other issue.

The figure reveals that, in general, respondents' attitudes on issue-pairs are broadly positively correlated, though this is somewhat more true for the treatment group than the control group (consistent with the modest positive effects documented in the main body of the paper for the constraint outcome). It is, however, notable that the correlations are all relatively low in absolute terms, with no issue pair having a correlation above .5. This implies that – even on issues that are reasonably closely related such as "Minimum Wage/Zero Hours" – a large fraction of respondents provide responses that are inconsistent with what we might expect if respondents were forming attitudes on traditional left-right ideological lines. This also implies that the null treatment effects documented in the paper are unlikely to be driven by ceiling effects, as it is clearly not the case that reason-giving fails to induce higher constraint because respondents' attitudes are already highly correlated across issues. In the "Sample One, Wave One" control group estimates, for instance, the correlation in issue positions ranges from -0.1 to 0.39 depending on the particular issue pair.

Polarization: Figure A7 presents the group-specific levels of polarization (measured using the mean absolute error of the survey responses on each item). There is clear evidence of cross-issue

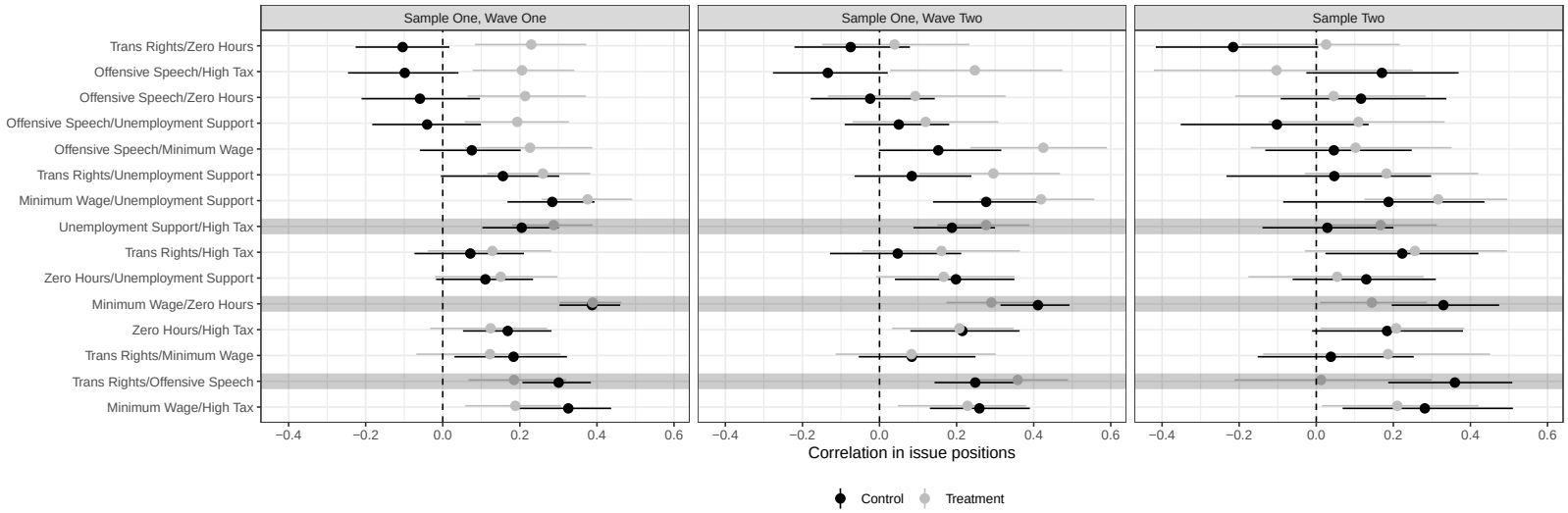


Figure A6: Treatment- and control-group issue-pair correlations

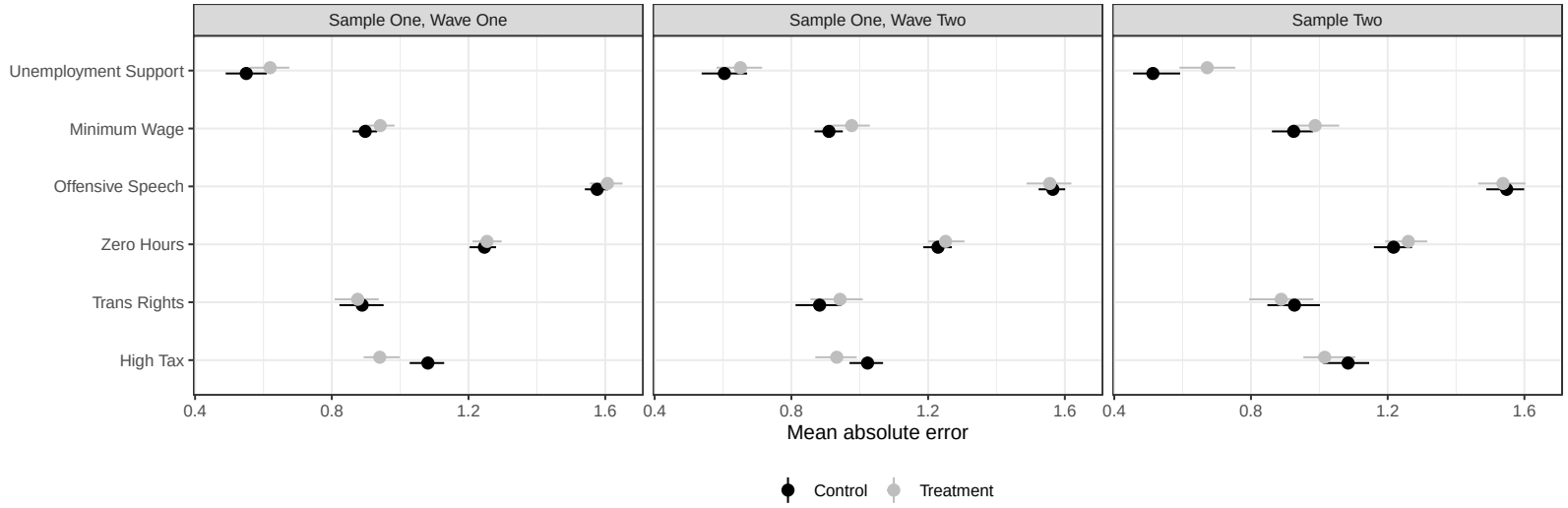


Figure A7: Mean absolute error (treatment and control)

heterogeneity in polarization, with responses to the "Offensive speech" issue more than twice as polarized as responses to the "Unemployment support" issue in both treatment and control groups. In addition, there is no evidence to suggest that the null effects reported in the paper are attributable to floor effects.

The MAE for the least divisive issue – unemployment support – is a little under 0.6, but even for this issue there are a large number of observations in the more extreme outcome categories. Figure A8 shows the raw response distribution for each policy, for both treatment and control groups, for the "Sample One, Wave One" respondents. As is clear from this figure, although the degree of polarization varies across issues, there is no issue where responses are so concentrated in a single category that reductions of polarization would be impossible. Together, this evidence again suggests that the null results presented in the paper are unlikely to be attributable to floor effects stemming from the polarization outcome measure.

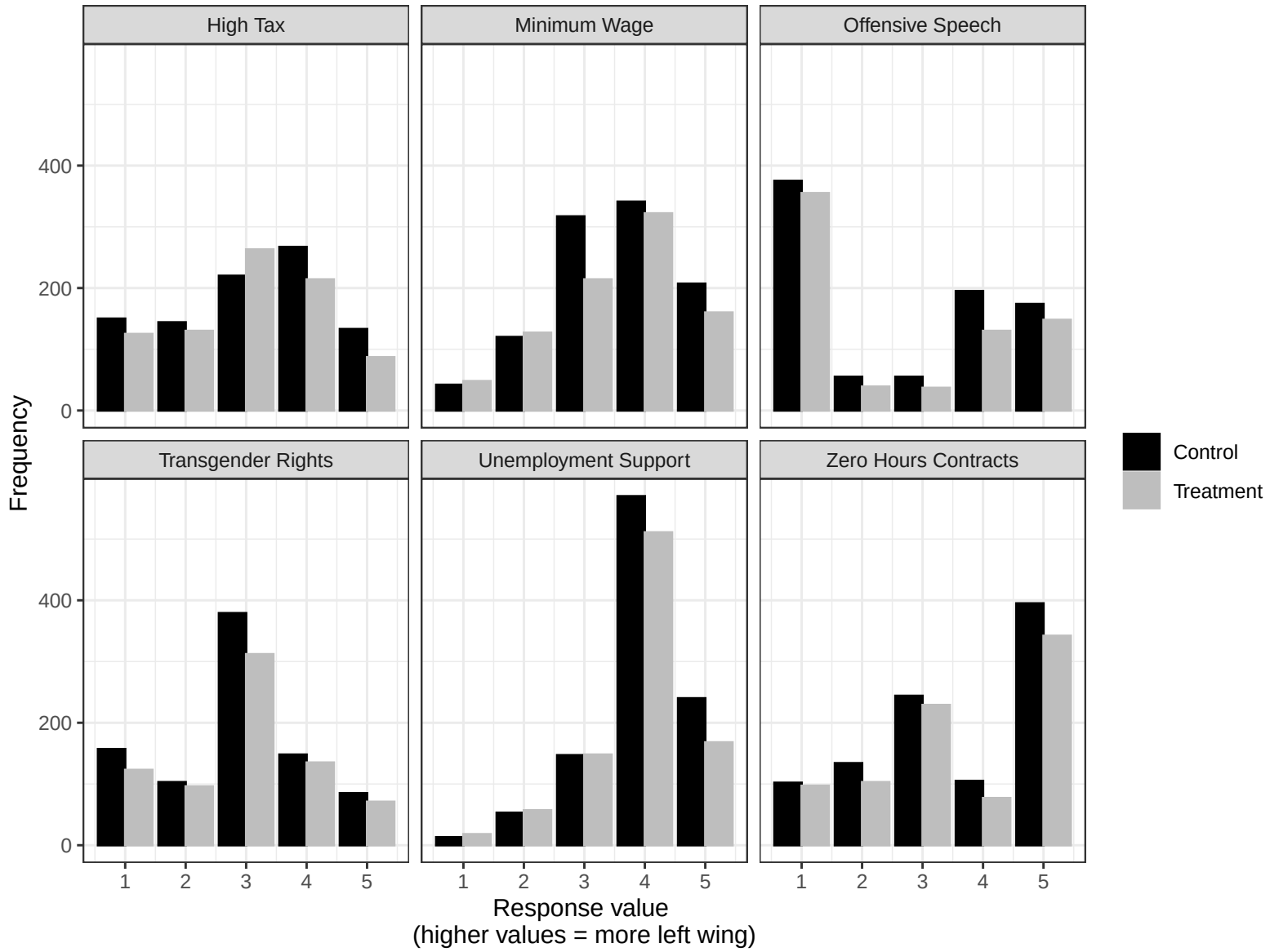


Figure A8: Raw outcome distributions (Sample One, Wave One)

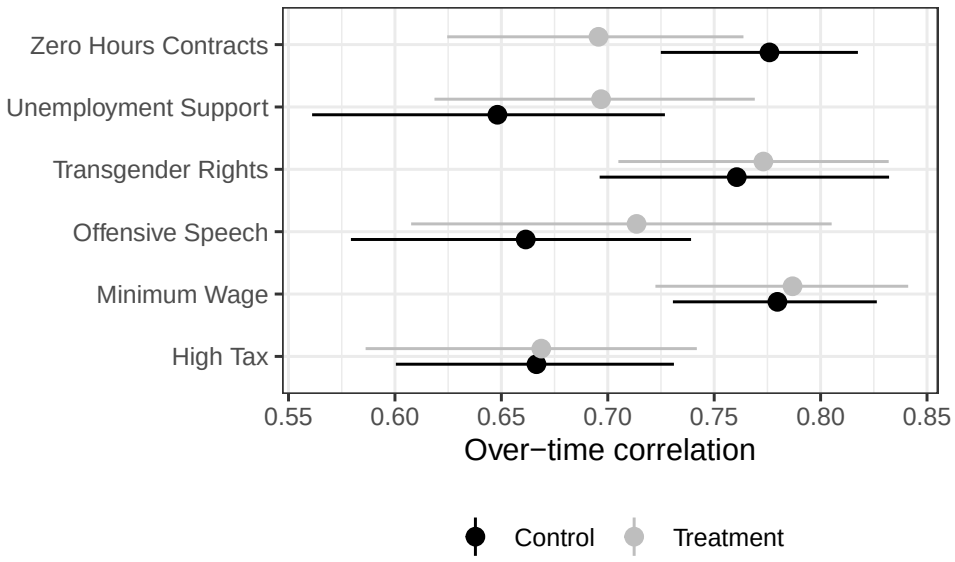


Figure A9: Treatment- and control-group over-time correlations

Stability: Figure A9 presents the group-specific levels of the stability outcome (the correlation in attitudes between survey waves). Across all six issues, the correlations are relatively high, with no issue-group combination having a correlation lower than .65. Correlations of this magnitude are comparable to levels of attitude stability reported elsewhere in the literature (Hanretty, Lauderdale and Vivyan, 2020), and although higher than the cross-issue correlations reported above, the correlations remain substantially below 1 implying that there is still room for the reason-giving treatment to take effect. In addition, looking across issues, there is no evidence that the null effects of the treatment are due to high baseline stability levels in the control group, as the magnitude of the estimated treatment effects does not appear to be related to the control group baseline levels.

F Alternative Measures of Polarization

The measurement strategy adopted in the main body of the text for the polarization outcome uses the difference in the mean absolute error of the survey responses on each policy item between the treatment and control groups. In this section, I consider two alternative measures of polarization: 1) the standard deviation of responses in each issue/treatment group; 2) the share of "extreme" responses (respondents selecting either option 1 or 5 in the ordered response scales) in each issue/treatment group.

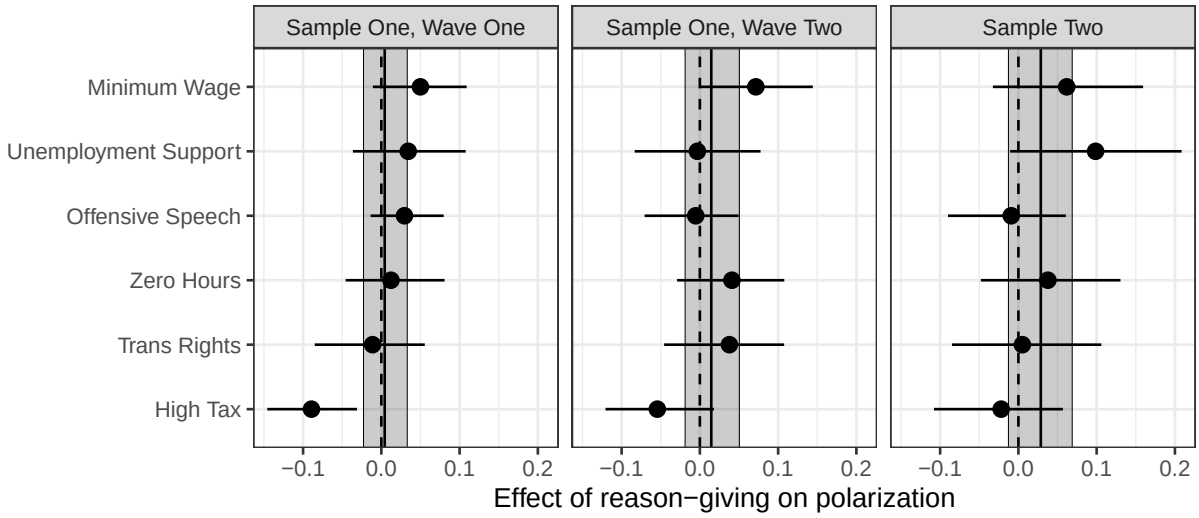


Figure A10: Effects of Reason-Giving on Polarization (Standard Deviation)

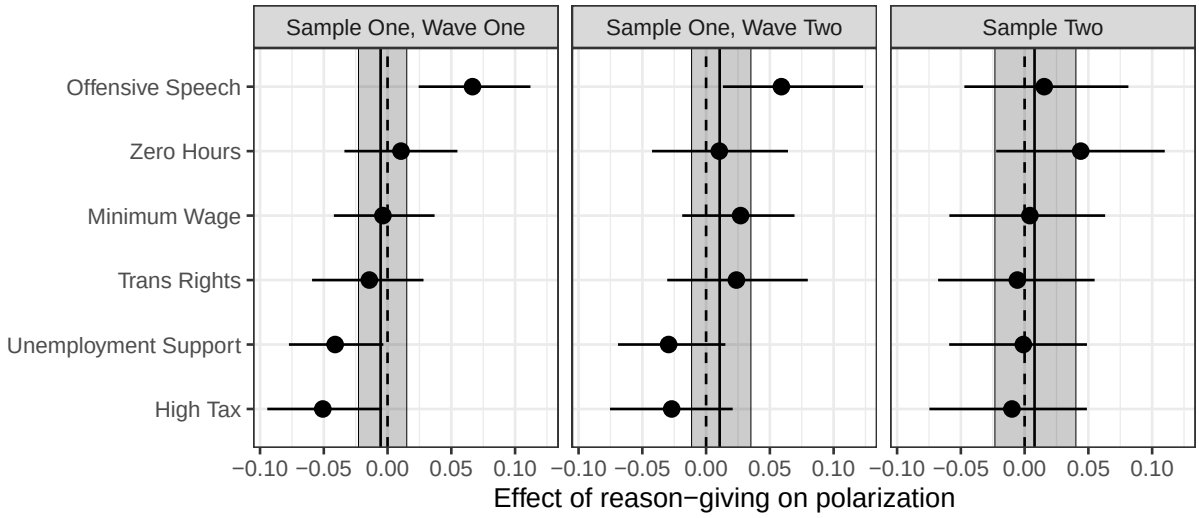


Figure A11: Effects of Reason-Giving on Polarization ("Extreme" responses)

Using these measures I then rerun the analyses depicted in figure 5 of the main body of the paper. Figure A10 depicts the estimated treatment effects using the standard deviation measure, and figure

A11 depicts the estimated treatment effects using the "extreme" responses measure. While there are some very modest differences at the issue level, the treatment effects calculated when averaging across issues are almost identical to those presented in the main body of the paper. This suggests that the null effects documented for polarization are not related to the particular metric of polarization I adopt.

G Treatment Effects on Left-Right Preferences

A plausible hypothesis is that – beyond any effects on stability, constraint or polarization – reason-giving might also affect respondents preferences on each of the issues included in the experiment. If we believed, for instance, that a given issue was more likely to result in a left-wing orientation after in-depth contemplation, but a more right-wing orientation on the basis of a "gut response", then reason-giving might result in respondents in the treatment group taking more left wing positions on that issue.

Figure A12 presents treatment effects for the average position taken on each issue. These coefficients come from bivariate linear regressions where I regressed the 5-point preference responses for each issue on a dummy for whether the respondent was in the treatment or control group. Positive coefficients represent issues where reason-giving respondents took more left-wing or socially-liberal stances on the issue, and negative coefficients correspond to issues where reason-giving respondents were more right-wing or socially-conservative than respondents in the control group. The vertical lines and confidence bands represent the effects of the reason-giving treatment on left-right preferences while averaging across issues, as estimated from a linear regression in which I stack the data for each issue and regress the preference variable on the treatment dummy and fixed effects for each issue (with standard errors clustered at the respondent level). For all models, I standardise the dependent variable to have mean zero and standard deviation one, such that the coefficients can be interpreted in standard deviations of the outcome.

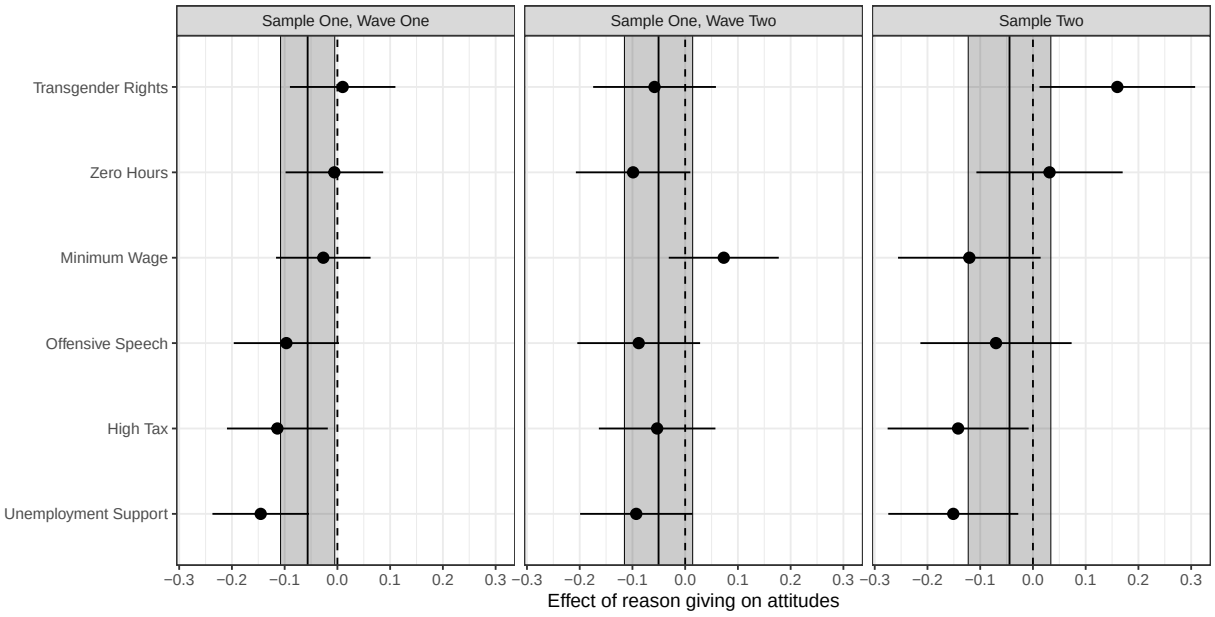


Figure A12: Effects of reason-giving on left-right position

The results show that, again, there are very minor effects of reason-giving on preferences. Across all three samples, there is a right-ward shift on average across issues for the reason-giving group of respondents, but this difference is very small in magnitude (about .05 of a standard deviation) and indistinguishable from zero except for the first sample of respondents in the first wave. At the level

of individual issues, there are also very small effects of reason-giving. There is some evidence that respondents shift further to the right on the issues of unemployment support and higher taxes for the wealthy, and somewhat to the left on the issue of transgender rights, but again these effects are small in magnitude and variable in significance. In sum, in addition to having limited effects on attitudinal constraint, polarization, or stability, reason-giving also largely fails to shift respondents towards either more liberal or more conservative issue stances on average.

H Reasons Given

What is the substantive content of the reasons given by respondents in the treatment group? Figure A13 depicts differences in word use across respondents with different policy preferences for each issue included in the experiment. The y-axis of these plots indicates the extent to which a given token (I use unigrams and bigrams here) is used more by one group than another.[12] Tokens higher on the y-axis (in blue) are used more by respondents who indicate agreement with the policy position given in the title of the relevant panel, while tokens lower on the y-axis (in red) are used more by respondents who indicate opposition to the policy position.

The figure reveals that the justifications that respondents provide contain language that is consistent with their expressed policy positions. For instance, respondents who are in favour of increasing the rate of income tax for higher income earners are much more likely to focus on the ability of those income earners to pay a higher rate of tax ("afford", "can_afford", "afford_pay"); more likely to characterise those subject to such taxes as "rich" while others are "poor"; and more likely to suggest that higher taxes have important societal benefits ("society", "contribute", "help", "services"). By contrast, those against tax increases on the rich give reasons which focus on issues of fairness ("fair", "high_enough", "work_hard") as well as on the possible consequences of higher taxes for economic activity (e.g. "incentive").

Similarly, proponents of increasing the minimum wage focus on issues relating to "cost", "poverty", "bills" and the standard of living, while opponents are much more likely to provide reasons focused on "companies", "businesses", "inflation", and the "market". For the offensive speech topic, those in favour of banning offensive speech are more likely to speak about the targets of such language ("racism", "race", "gender") and the consequences of offensive language ("speech_can", "behaviour", "abuse"), while those in opposition tend to focus on "free_speech", and the idea that people are too easily offended.

Very similar patterns can be seen across the other issues in the experiment, with distinctive words arising between groups in each case. Taken together, these differences suggest that respondents were engaging with the reason-giving treatment in the experiment, as people provided justifications that were substantively related to the policy preferences that they subsequently went on to express.

[12] In particular, I use the Z-score of the log-odds-ratio for each word, as described in Monroe, Colaresi and Quinn (2008).

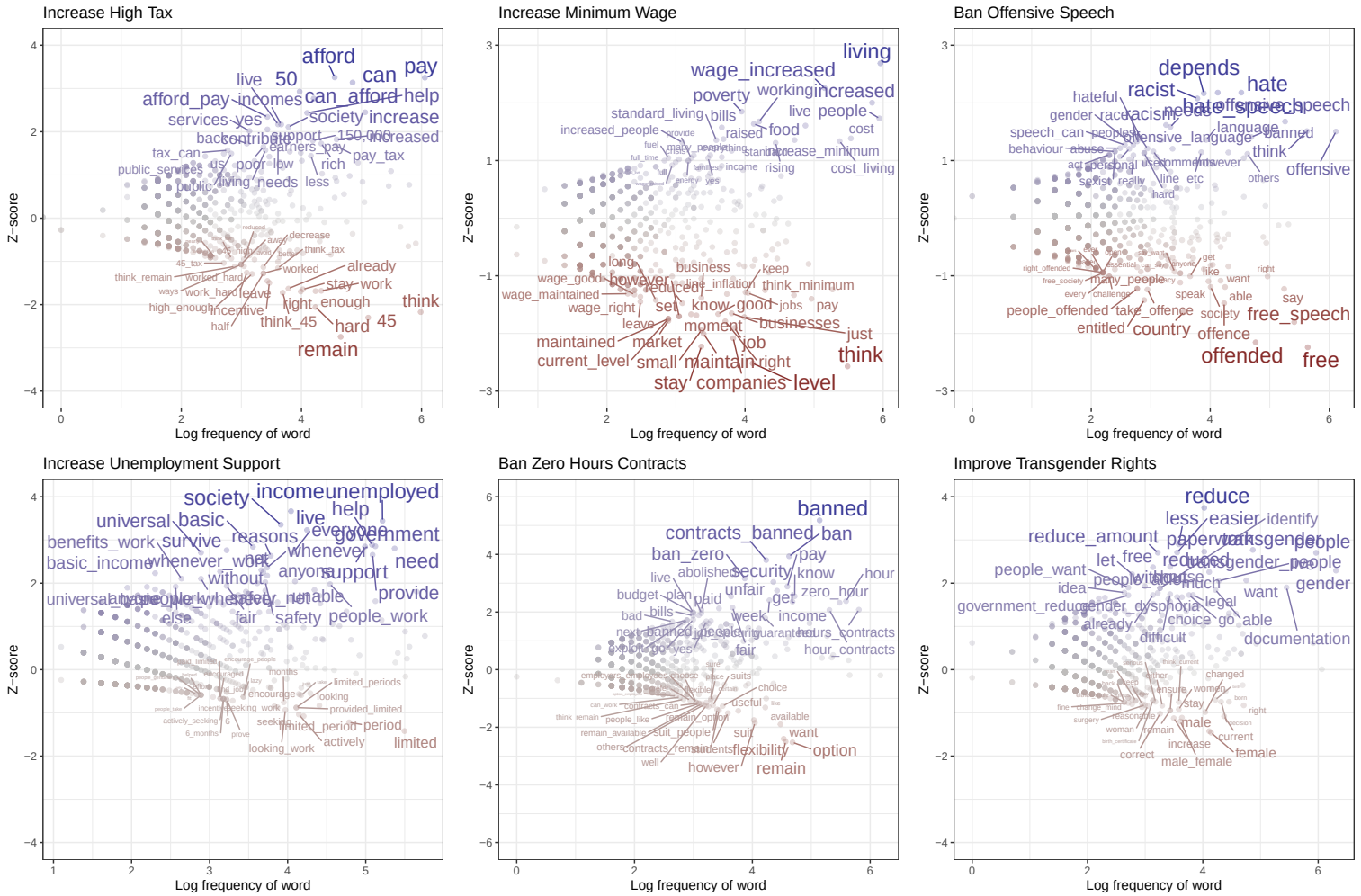


Figure A13: Distinctive token use by issue position

The figure shows the tokens that are most strongly associated with survey respondents on each side of the 6 issues included in the experiment. The y-axis plots the Z-score of the log-odds ratio for a given word, a quantity which measures the difference in token usage between respondents in favour of the issue position in the title of each panel (in blue, higher on the plot) and respondents against the issue position (in red, lower on the plot). The x-axis plots the (logged) token use in the corpus as a whole.