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 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 constraint, 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

9972 words

*This version: June 03, 2024. 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 deliberative democrats have argued that the public exchange of reasons between individuals can "change minds and transform opinions" (Chambers, 2003, 318; see also Dryzek, 2002; Cohen, 2005; Thompson, 2008; Mutz, 2008; Gutmann and Thompson, 2009). In addition to the attitudinal effects of inter-personal deliberation, scholars have also speculated that justifying one's political attitudes may also induce a greater degree of "internal-reflective" (Goodin, 2000) deliberation and introspection which, in turn, might affect the content of political attitudes. For instance, deliberative processes in which the weighing of reasons "take[s] place within the head of each individual" (Goodin, 2000, 82) are thought to affect "how we decide what position to take" (Goodin and Niemeyer, 2003, 629). Similarly, Cohen (2007, 228) suggests that "the practice of defending proposals with reasons may change my preferences." However, while voters are able to provide substantive reasons in defense of their preferences (Colombo, 2018, 2021), very little existing research evaluates whether and how introspective 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).

How might reason-giving affect attitudes? On the one hand, models from political behaviour hold that voters' reasons play a causal role in the construction of their attitudes (e.g. Zaller, 1992; Zaller and Feldman, 1992). *Reasons-as-causes* models of this form highlight that, by encouraging them to introspect about their preferences, reason-giving might change the set of reasons that voters consider, and thereby affect the attitudes they express. This model of attitude formation leads to three specific expectations. First, reason-giving might increase the *temporal stability* of voters' attitudes by reducing judgments made on the basis of idiosyncratic, "top of the head", considerations. Second, reason-giving might increase the *ideological constraint* of voters' attitudes by highlighting substantive connections across issues. Third, reason-giving might reduce issue-based *polarization* if deeper introspection reduces the variance of voters' expressed attitudes, or if it encourages voters to consider arguments that support positions other than their own, thereby reducing voters' attitudinal extremity. On the other hand, models from social and political psychology 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). These *reasons-as-rationalizations* models imply that, if reasons are used to rationalise (rather than cause) political beliefs, there are few mechanisms through which introspecting about 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 implement an experimental design in which 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. I evaluate the effects of reason-giving by measuring differences between treatment and control with respect to constraint (the correlation between respondents' positions on different issues), stability (the correlation of respondents' positions across survey waves), and polarization (the disagreement between respondents' positions on a given issue).

Fielding this pre-registered experiment in a new two-wave panel survey of more than 4,000 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. 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, this paper helps to open up the "black box of deliberation" (Mutz, 2008, 531) by demonstrating that one particularly important component of deliberative practice – the articulation of reasons – has essentially no effect on the attitudes that voters express. As others have noted, developing empirical evidence on the consequences of specific components of deliberation is an important endeavour, which "greatly enhance the capacity of deliberative theory to contribute to democratic society" (Mutz, 2008, 531).

In particular, the results here speak to the relative efficacy of public versus private deliberation. For many deliberative democrats, inter-personal exchange is critical for realising the benefits of deliberation (e.g. Dryzek, 2002; Rawls, 1997; Gutmann and Thompson, 2009; Cohen, 2005). For others, by contrast, deliberation between people is just 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). One motivating factor for studying the internal-reflective form of deliberation is that inter-personal exchanges of the type that occur in citizen assemblies, citizen juries, and other deliberative experiences are hard to scale to large populations. If the types of attitudinal change 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. As Goodin (2000, 84) suggests, internal-reflective deliberation may therefore "relieve many of the burdens plaguing external-collective deliberation in modern mass societies." Empirical studies of the effects of internal-reflective deliberation are, however, rare.[1] My results suggest that – at least with respect to political attitudes – solitary reason-giving does not have effects equivalent to public deliberation. As a consequence, the benefits ascribed to broader deliberative practices must therefore be generated by components of deliberation other than the private, internal form of reason-giving studied here.

Finally, testing the expectations generated by the *reasons-as-causes* model is important because they imply an optimistic view of voters' capacity to hold well-structured political preferences. Decades of survey research has painted a pessimistic picture about the stability, constraint, and polarization of voters' attitudes (e.g. Converse, 1964; Achen, 1975; Zaller and Feldman, 1992; Ansolabehere, Rodden and Snyder, 2008; Freeder, Lenz and Turney, 2019; Abramowitz and Saunders, 2008; Fiorina and Abrams, 2008; Mason, 2015). Deficiencies in these dimensions pose an obvious a normative threat: if voters hold unstable, incoherent and extreme policy views, then their ability to hold politicians accountable is consequently diminished (see, e.g., Achen and Bartels, 2017, 306). The hopeful suggestion generated by the *reason-as-causes* model is that the quality of voters' attitudes might increase if only voters could be induced to "think harder" about their political opinions, a suggestion buttressed by evidence from social psychology that shows greater introspection can indeed lead to more stable, more coherent, and less polarized attitudes in other domains (e.g. Petty and Briñol, 2011; Tesser, 1978; Wilson et al., 1993; Wilson, Kraft and Dunn, 1989; Dijksterhuis, 2004). However, I show that increased cognitive effort does not result in such salutary effects for three important measures of *political* attitude quality (Price and Neijens, 1997). Notwithstanding any intrinsic value of reason-giving, the results here therefore imply that more introspective and reason-based processing of political issues is unlikely to act as a panacea to the problem of low-quality democratic attitudes.

[1] Though see Minozzi et al. (2023) for an important recent example.

2 Reason-Giving and Political Attitudes

Reason-Giving and Normative Democratic Theory

In liberal democratic theory, 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; Chambers, 2010; Habermas, 2015). The centrality of the "reason-giving requirement" (Gutmann and Thompson, 2009, 24) in deliberative democracy, for instance, 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 this intrinsic virtue, however, deliberation is also thought to "change minds and transform opinions" (Chambers, 2003, 318), and reason-giving is seen as a central mechanism through which such effects might operate.

For many scholars, deliberation is thought to affect preferences through the public and social *exchange* of views between different people (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, any process in which voters are induced to be more reflective about their political positions can therefore be considered deliberative, regardless of whether such processes include the public exchange of reasons (Goodin and Niemeyer, 2003, 629). Public reason-giving might be one way of inducing internal-reflection, but it is not the only way. As Goodin (2000, 95) argues, "sometimes 'answering to oneself' might suffice." The key insight from this work is that solitary reason-giving might affect preferences via the internal process of introspection it engenders in voters, rather than through a public process in which voters exchange reasons with one another.[2]

However, beyond the broad hypothesis that reason-giving might affect attitudes, work in normative theory provides few operationalizable predictions about the effects of reason-giving on specific attitudinal outcomes. In part this is because these accounts do not (and were not designed to) clearly

[2] Similar arguments can be found in Cohen (2005, 349) and Bortolotti (2009, 642).

articulate the psychological mechanisms through which reason-giving might lead to attitude change. In the next section, I contrast two models of attitude formation which take different perspectives on the role that reasons play when voters think about politics and which generate different expectations for the effects of reason-giving on 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 on a given issue changes, then the voter's opinion on that issue may also change. The idea that attitudes are causally determined by aggregating across reasons is shared by many accounts,[3] but the most prominent example of such an argument comes from Zaller (1992).[4] 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 critical assumption here is the idea that voters draw a *sample* of reasons each time they are required to produce a political opinion, and it is from this sample they then construct their attitudes. This assumption drives many of the predictions derived below, as the additional cognitive effort that reason-giving induces is expected to affect attitudes by changing the sample of reasons that voters consider.

By contrast, *reasons-as-rationalizations* models see political attitudes as deriving from fast and intuitive processing in which explicit reasoning plays a very limited role. Lodge and Taber (2013), for

[3] This argument appears in social psychology (e.g. Azjen, 1980), survey research (e.g. Tourangeau and Rasinski, 1988), as well as in the expectancy-value framework that underpins the literature on framing effects (e.g. Nelson, Oxley and Clawson, 1997; Chong and Druckman, 2007).

[4] See also Zaller and Feldman (1992).

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 spontaneous and affect-driven processes which are entirely unrelated to the evaluation of specific reasons. The idea that people will provide evaluations without engaging in a cognitive reasoning process is also common in both social (e.g. Mercier and Sperber, 2018) and moral (e.g. 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 valence of the 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-rationalizations* 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.

In general, both approaches understand attitude formation as a fast and constructive process in which attitudes are generated at the moment of response, rather than existing as a fixed point in voters' minds. In this sense, both approaches 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). Where the approaches disagree, however, is in the mechanisms by which the process of attitude construction occurs. The critical distinction is that while *reason-as-causes* models suggest a cognitive process in which reasons are aggregated to form attitudes, *reasons-as-rationalizations* models suggest an affect-based process in which attitudes are adopted spontaneously without any evaluation of specific reasons. These differences in perspective about the internal workings of the attitude formation process are relevant because they imply very different predictions for the distribution of attitudes when voters are induced to engage in slower, more effortful contemplation.

What do these models predict for the effects of reason-giving on political attitudes? First, *reason-*

as-causes models suggest that reason-giving might affect 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 (Zaller, 1992, 86).[5] 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. 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.

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 explicitly stat-

[5] 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 $\sigma^2_{i,p}$ 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).

ing justifications might also make voters aware of conceptual links across different issues, thus inducing them to express more correlated attitudes (e.g. Keating and Bergan, 2017). 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 reasons 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. I conceptualize polarization as the extent of disagreement between voters' issue positions on a given issue. Decreases in polarization 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. 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).

The common logic underpinning the expectations from the *reasons-as-causes* model is that reason-giving might change the set of considerations that voters use to construct their attitudes. These expectations are substantively important because they imply that three key properties of attitude quality might be improved simply by voters exerting by a greater degree of cognitive effort. What, then, does the *reasons-as-rationalizations* 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 only 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.

Nevertheless, the *reason-as-rationalization* perspective has nuanced predictions for the effect of reason-giving on polarization. If people reason in a biased manner, their initial affective reactions might be further reinforced by the accumulation of reasons that align with that response. As a consequence, such voters might develop greater *confidence* in the attitudes they express, as the reasons drawn to mind could justify and validate their initial intuitive responses. This type of biased processing may also lead voters to adopt more *extreme* views. For instance, if the reasons a voter recalls all align with the particular side of a debate to which the voter is intuitively attracted, considering those reasons might prompt them to reconsider their initial response as being too moderate, and encourage them to take a more extreme position on that issue (Tesser, 1978, 310). Under the *reasons-as-rationalizations* model, then, reason-giving should be expected to have either null effects on aggregate polarization (if reasoning only increases attitudinal confidence), or positive effects on polarization (if reasoning increases attitudinal extremity). Importantly, both of these predictions differ from those derived from the *reasons-as-causes* model, which implies that introspective reason-giving will reduce attitudinal polarization.

In the context of the typical survey response, both models suggest that voters make fast, and largely unconscious, judgements. For *reasons-as-causes* models these judgements arise via a fast and shallow sampling of reasons which are then used to determine their choices, while for *reasons-as-*

rationalizations models they stem from intuitive, affect-based, and spontaneous responses. As a consequence, both models are consistent with commonly observed response patterns in many political surveys in which voters' attitudes are marked by low levels of constraint and stability, and high levels of polarization. However, contrasting predictions of these models arise when considering the expected effects of increased effortful thinking: where the *reasons-as-causes* approach assumes that such effort will change the set of considerations brought to mind and therefore the resulting attitudes that voters express, the *reasons-as-rationalizations* approach assumes that effortful thinking will produce reasons that justify the initial affective response of the voter and will have few consequences for expressed attitudes.

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, Kraft and Dunn, 1989; 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, these studies do not directly address reason-giving as a specific mechanism for attitude change. Moreover, many of these papers focus on consumers' choices, which limits the degree to which they are informative about political attitudes.[6]

In political science, 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), 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

[6] Though see Wilson, Kraft and Dunn (1989, study 2).

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 studies 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).

In a recent study, Minozzi et al. (2023) focus specifically on evaluating the separate effects of public and private deliberation on a range of outcomes, such as knowledge gains, emotional reactions, and civic attitudes. Consistent with the results I present below, they find only limited effects of individual deliberation. However, the treatment that Minozzi et al. (2023) employed differs in important ways from the the treatment I introduce below, most notably in that it did not require participants to engage in reason-giving. This, in addition to the fact that Minozzi et al. (2023) do not study the effects of individual deliberation on the quality of voters' attitudes, suggests that further research into the specific effects of introspective reason-giving is warranted.[7]

The study that comes closest to evaluating the effects of reason-giving on attitudes is by Zaller and Feldman (1992)[8] 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

[7] The design used in Minozzi et al. (2023) is also only powered to detect large treatment effects of individual deliberation (see appendix E of their study), something that also warrants further research.

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

(insignificantly) *lower* in the stop-and-think condition than in the control condition (Zaller and Feldman, 1992, 605).

However, 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,[9] 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. 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."

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].

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

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

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 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.

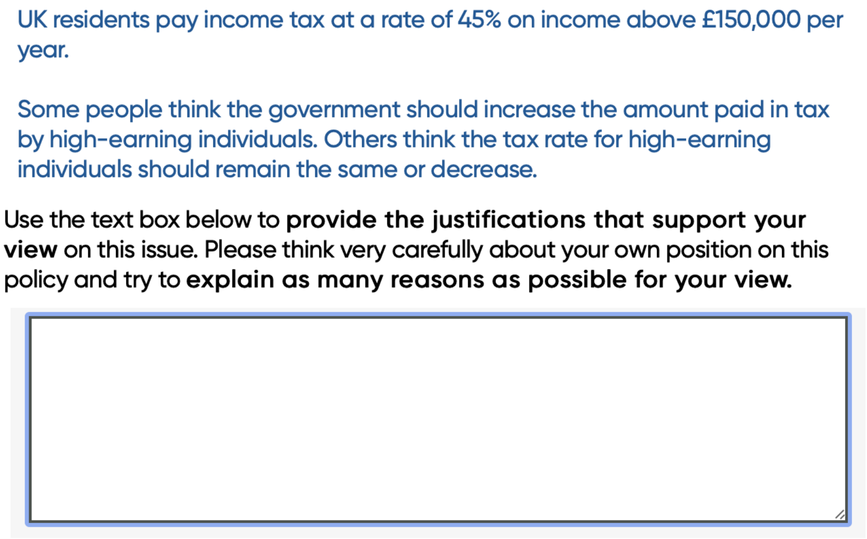
Policy Areas

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, attitudes on which are thought to be structured by different types of cognitive processes (Carmines and Stimson, 1980). 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.

The design is only sufficiently powered to detect relatively large treatment effects at the level of individual issues (see appendix section B). 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 and to provoke the type of introspection that the *reasons-as-causes* model predicts will be consequential. 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 pro-



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.

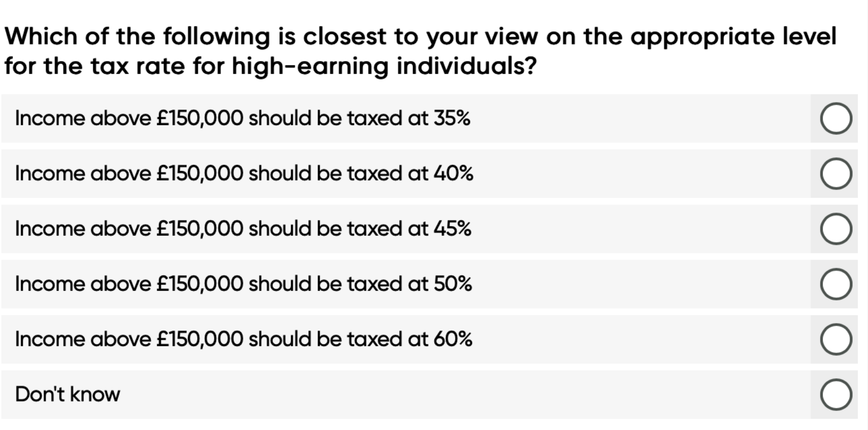
Use the text box below to **provide the justifications that support your view** on this issue. Please think very carefully about your own position on this policy and try to **explain as many reasons as possible for your view.**

Figure 1: Reason-giving prompt

vides 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).

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 introduction to the policy issue (the blue text visible in figure 1) and the issue-position prompt in figure 2, but were were not asked to provide reasons supporting their attitudes. The full text of both prompts for each of the six issues included in the experiment is given in appendix A.



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 ○

Figure 2: Issue position prompt

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*). 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.[10]

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

[10] Averaging across issues in the first wave of the survey, 14% of responses were "Don't know" responses. 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).

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 *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 issue pairs as reason-giving might affect constraint by making respondents aware of common justifications that apply across 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. $MAE^{D=0}_{\text{HighTax}}$).[11] 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. 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.

The MAE statistic reflects the conceptualization of polarization as the extent of disagreement between voters' issue positions on a given issue. I focus on this measure, rather than the proportion

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

$$MAE_k^{D=d} = \frac{1}{\sum w_i} \sum_{i=1}^{N_{D=d}} w_i|\mu_k^{D=d} - X_i^k|$$

where X_i^k 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.

of voters who adopt "extreme" issues positions, because it is possible for a decrease in issue-based disagreement to occur in the absence of voters adopting uniformly more moderate positions. For instance, if reason-giving were to shift a large group of voters with moderate positions a little to the right, and a small group of very right-wing voters a little to the left, the result would be a decrease in polarization (disagreement between the groups would have declined) but an increase in the average extremity of voters' attitudes (the median voter would have more right-wing attitudes than previously). Accordingly, I focus on measuring the effects of reason-giving on the degree of disagreement among voters on a given issue, rather than the share of voters who adopt extreme issue position. However, in supplementary analyses in appendix section F I demonstrate that the results are unaffected by using alternative measurement strategies for polarization.

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.

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.

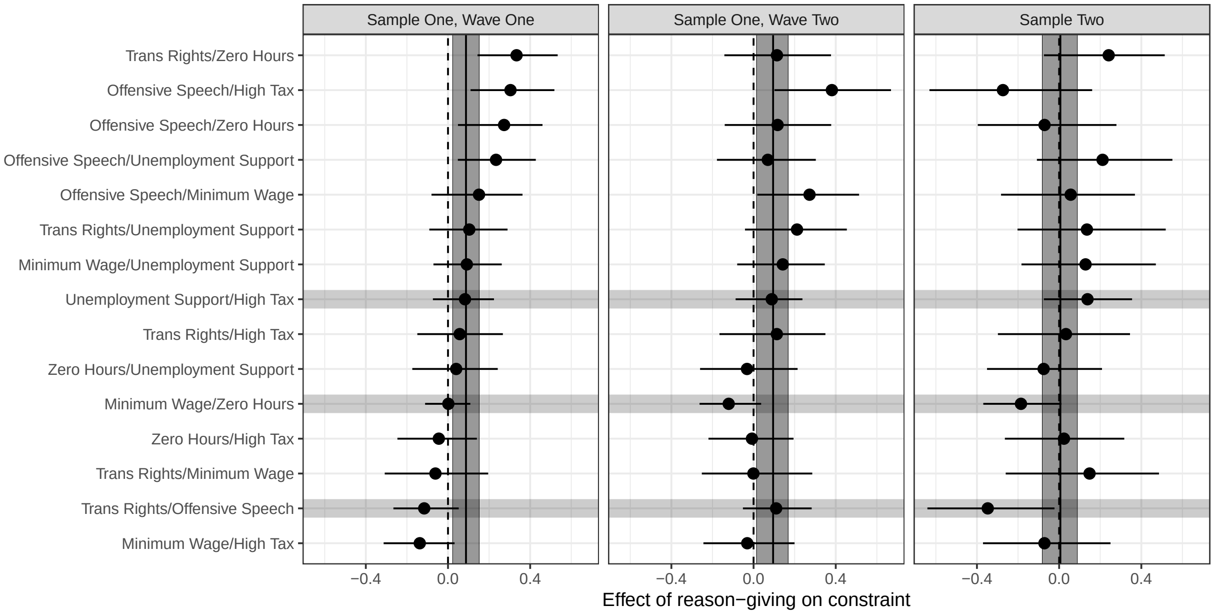


Figure 3: Effects of Reason-Giving on Constraint

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. 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. For 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 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 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,

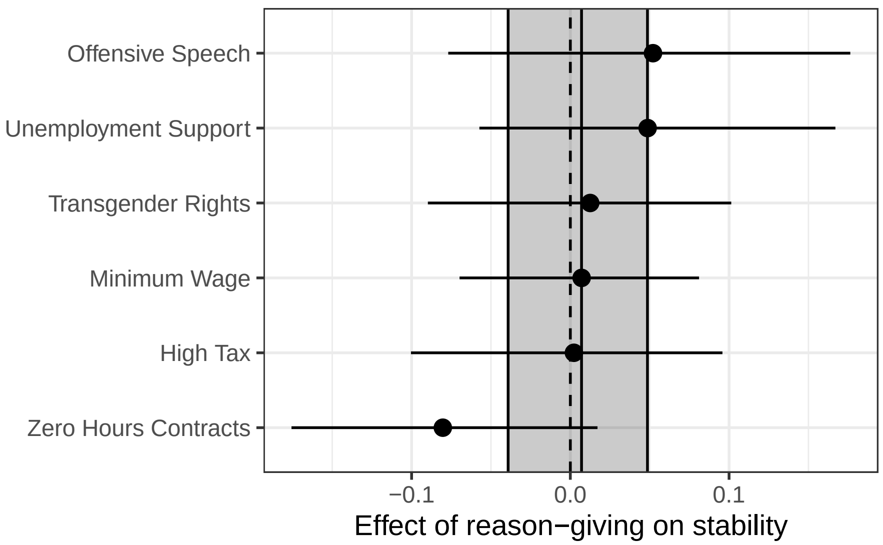


Figure 4: Effects of Reason-Giving on Stability

0.049]). 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-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.

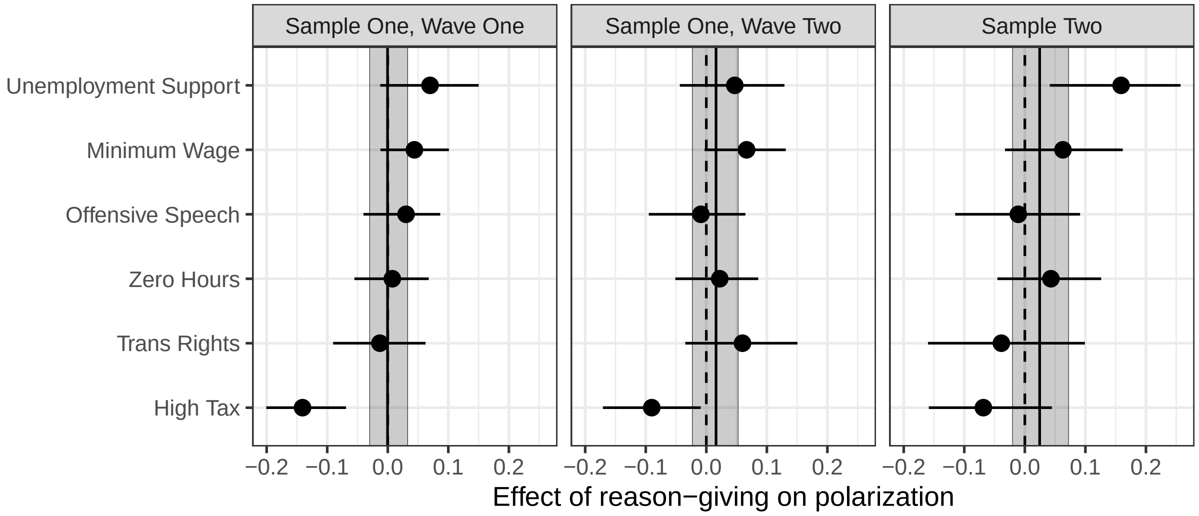


Figure 5: Effects of Reason-Giving on Polarization

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 [-0.029, 0.033]) and wave two (0.016 [-0.024, 0.051]) of the survey. The same is true for the second sample of respondents where the estimated treatment effect is 0.024 [-0.019, 0.073]. 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

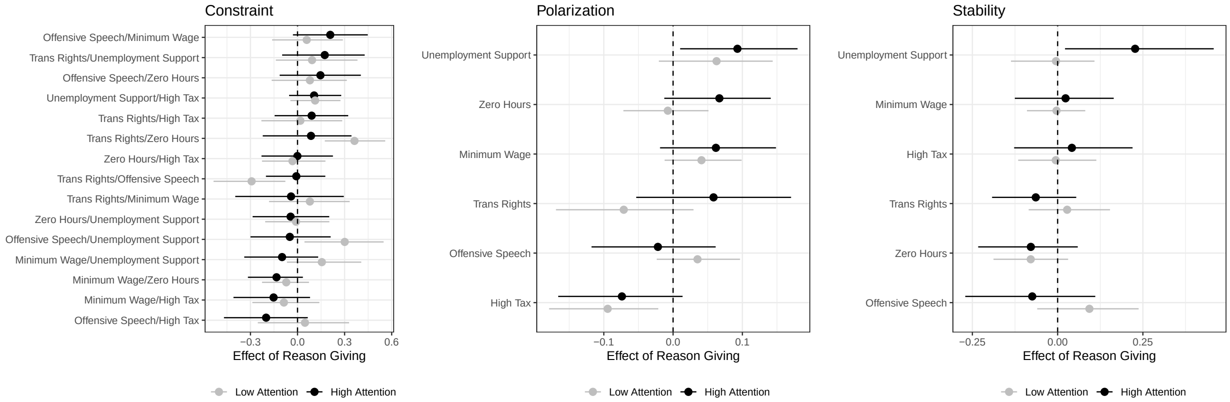


Figure 6: Conditional Issue-Level Treatment Effects by Political Attention

attention.[12]

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 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 appendix figure A16, in analyses that were not pre-registered, I explore whether the average (i.e. across issues) effect of the reason-giving treatment on each outcome varies across different groupings of respondents, determined by age, gender, education, political attention, and past vote in the 2016 Brexit referendum and the 2019 general election. I find very little evidence of systematic heterogeneity across these groups. 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.

6 Threats to Inference

One potential objection is that if the reason-giving treatment did not provoke respondents to think more deeply about their attitudes, then the null effects reported above might be attributable to the

[12] As pre-registered, I divide respondents according to whether they are above or below the median on this 11-point variable. To maximise power, I pool responses from the first and second samples for this analysis.

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 [66.9, 71.1] 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. Appendix section C also includes further analyses which demonstrate that these differences in average engagement are not driven by any particular subset of treatment group respondents and that results are not affected by subsetting to units who engaged with the reason-giving treatment at greater length.

Second, the content of the reasons provided by respondents in the treatment group suggests a high degree of engagement with the underlying issues. The median length of responses to the open-ended reason-giving prompt was between 15 and 22 words, depending on the 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 treatment respondents provided 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 A15). This again suggests that people were following the instructions in the prompt and actively considering the reasons that lie behind their political beliefs. Overall, it is therefore unlikely that the reason-giving prompt failed to compel respondents to canvas their minds for salient considerations.

An additional threat to inference is that the reason-giving treatment requires 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. Given that the treatment requires respondents to engage in an effortful political-reasoning task, it is plausible that the estimated treatment effects could 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.

Readers may also be concerned that the time between the first and survey waves is longer than is typically the case in survey experiments. This could potentially bias the stability analysis towards a null result as the effect of reason-giving has a reasonably long period of time to dissipate. While definitively ruling out this explanation would require replicating the experiment across shorter time horizons, it is worth noting that the design here differs from survey experiments which seek to measure the effect of an information-provision treatment in wave one on attitudinal responses in wave two. Here, the reason-giving treatment is repeated for all treatment-group respondents in the second wave of the survey. This repeated exposure reinforces the treatment strength and makes it more likely for any stability-inducing effects of reason-giving to manifest, despite the somewhat longer time period between survey waves.

Finally, readers might wonder whether reason-giving has effects on other properties of attitudes. One obvious hypothesis is that reason-giving respondents might provide responses that are systematically further to the left or the right on a given issue. In appendix G, I show that although some differences appear on individual 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 core contribution of this paper is to show that reason-giving does not, in isolation, have the salutary effects on political attitudes predicted by the *reasons-as-causes* model and hoped for by proponents of deliberative democracy. Many scholars are optimistic that deliberation can profoundly affect the quality of the attitudes that voters hold, and recent work has explored the potential for "reflective, intrapersonal, and private" thought (Minozzi et al., 2023, 2) to act as a mechanism for delivering the benefits of deliberation. Indeed, for some, "internal-reflective" deliberation "might even be a more important part of the process than the dialogic and discursive element" of deliberation (Goodin and Niemeyer, 2003, 628). I argued that the *reasons-as-causes* model helps to clarify how such introspection, as induced by reason-giving, might lead to specific effects on a series of important measures of attitudinal quality. The normative importance of these expectations is clear: if greater cognitive effort could help to save voters from having "vague, uninformed, or incoherent" (Achen and Bartels, 2017, 108) attitudes, then the prospects of strengthened democratic accountability would be consequently enhanced. The null results presented here, however, suggest that whatever weaknesses exist in the political attitudes of the public, inducing voters to devote more cognitive effort to the reasons that underpin their attitudes is insufficient for improving the quality of those attitudes.

The findings here do not, of course, undermine the claim that deliberation, *in toto*, might have beneficial effects on democratic attitudes. I took seriously calls to investigate "important, specifiable, and falsifiable" claims in deliberative democratic theory (Mutz, 2008, 521) by focusing attention on understanding the specific effects of introspective reason-giving, but there are alternative mechanisms by which deliberation might affect attitudes. First, the treatment employed here aimed to solicit "internal-reflective" reasoning, but it might miss potentially important effects stemming from the *public* exchange of political reasons (e.g. Rawls, 1997; Mercier and Sperber, 2018). Second, deliberation might also expose voters to new information about different policy options, and that information might affect expressed preferences. Future work should therefore investigate whether different *types* of reason-giving and different elements of deliberation have effects on political attitudes, and under which conditions.

These results also have important implications for existing models of attitude formation. In particular, the results contrast with the expectations generated from the model presented in Zaller (1992). The critical assumption of that model is that, at the point of attitude construction, voters sample considerations over which they then aggregate to form attitudes. The additional cognitive effort induced by reason-giving is therefore expected to affect attitudes by changing the sample of reasons that voters consider. However, voters' issue attitudes appear to be largely insensitive to the amount of introspective reasoning in which they engage. These null effects therefore cast doubt on the idea that voters construct attitudes via such a cognitively-based, consideration-sampling process and instead are more consistent with the idea that voters' attitudes primarily form via instinctive, affect-driven reactions (e.g. Lodge and Taber, 2013).

It is important to note, however, that Zaller's consideration-sampling assumption is analytically separable from the assumption that reasons play a causal role in the construction of voters' attitudes. For instance, it is possible that voters form attitudes by averaging over attitude-relevant considerations but do not sample different considerations in each instance. If this is the case, then even though reasons are playing a causal role in attitude formation, we would not expect reason-giving to have large effects, as the reason-averaging process would make use of the same considerations at all points in time. Therefore, while the results here contrast with the predictions of the Zallarian consideration-sampling view, they do not necessarily represent strong evidence against *reasons-as-causes* models as a whole. Nevertheless, given that Zaller's (1992) consideration-sampling logic is used as a foundation for many arguments in political behaviour (e.g. Bullock and Lenz, 2019, 327; Freeder, Lenz and Turney, 2019, 288; Baccini and Leemann, 2021, 471), demonstrating that the predictions of that model are not supported by this experiment is an important contribution to the debate over the psychological mechanisms that underpin the expression of political attitudes.

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 (Price and Neijens, 1997). However, these properties do not represent all the potentially relevant outcomes which might be affected by reason-giving. An interesting prediction of the *reasons-as-rationalizations* model is that by engaging in a process of reason-giving, voters will draw to mind considerations that buttress their intuitively formed attitudes, thus increasing the confidence with which they express those attitudes. One important omission here is therefore the absence of data on the *strength* of voters' attitudes, and further research might profitably explore whether reason-giving has such effects on opinion strength. Similarly, another interesting avenue would be to explore whether the exchange of reasons between voters of different political opinions might help to decrease hostility across lines of political disagreement.

Finally, my results contrast with a well-established literature in social psychology which finds finds that asking people to explain the reasons for their attitudes can change the attitudes that they express (e.g. Wilson, Kraft and Dunn, 1989; Wilson and Schooler, 1991; Wilson et al., 1993; Dijksterhuis, 2004; Simonson, 1989; Hsee, 1999). In most cases, this research focuses on reason-giving in non-political settings, which provokes the question of whether there is something distinct about the process of reasoning about politics that prevents introspection from having the effects that are apparent elsewhere. One possibility is that people are less informed or knowledgable about their political attitudes, and so the quality of their introspective reasoning is lower than for affairs with which they are more familiar. Another possibility is that the affective reactions that people experience when thinking about politics are stronger than in other domains, and thus subsequent reasoning is more likely to be biased in the direction of their initial response. Answering these questions is beyond the scope of this paper, but exploring why there are differences in reason-giving effects across settings would be another interesting direction for future research.

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* .

Baccini, Leonardo and Lucas Leemann. 2021. "Do natural disasters help the environment? How voters respond and what that means." *Political Science Research and Methods* 9(3):468–484.

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.

Bullock, John G and Gabriel Lenz. 2019. "Partisan bias in surveys." *Annual Review of Political Science* 22:325–342.

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.

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.

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.

Keating, David M and Daniel E Bergan. 2017. "Mapping political attitudes: The impact of concept mapping on ideological constraint." *Communication Studies* 68(4):439–454.

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"?

Mason, Lilliana. 2015. ""I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization." *American Journal of Political Science* 59(1):128–145.

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

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.

Tourangeau, Roger and Kenneth A Rasinski. 1988. "Cognitive processes underlying context effects in attitude measurement." *Psychological bulletin* 103(3):299.

Wilson, Timothy D, Dolores Kraft and Dana S Dunn. 1989. "The disruptive effects of explaining attitudes: The moderating effect of knowledge about the attitude object." *Journal of Experimental Social Psychology* 25(5):379–400.

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

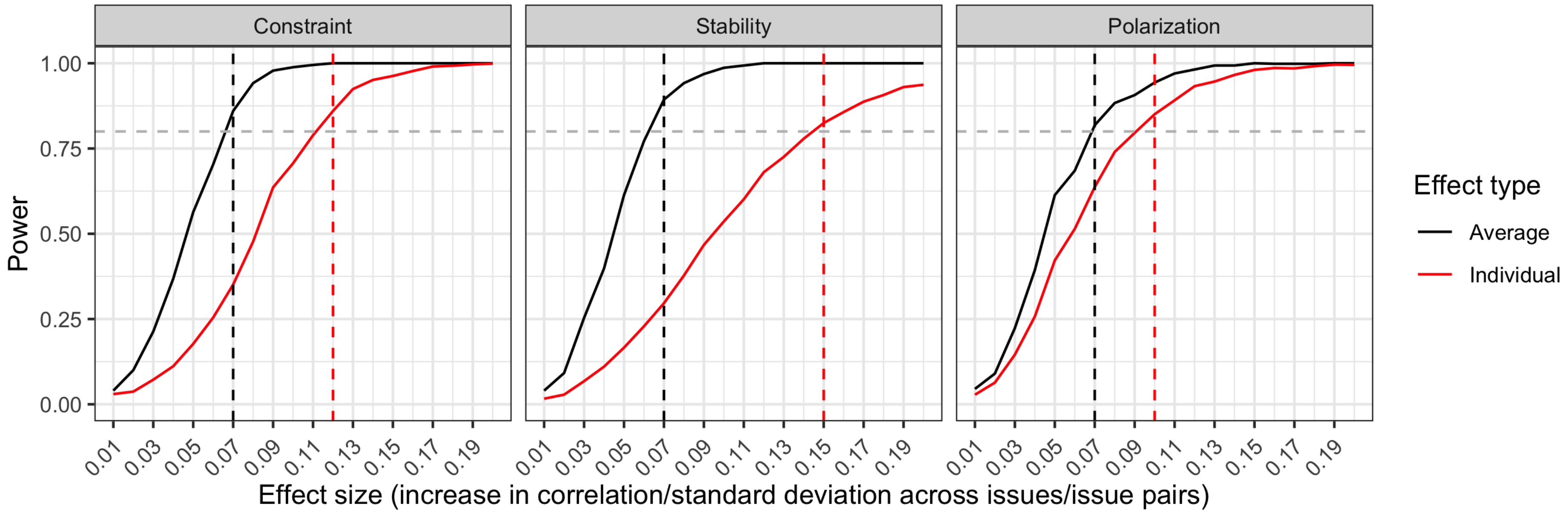
Figure A1 shows the results of a power analysis for the quantities of interest described in section 4 of the paper. To construct the power analysis, I simulated the data collection process for a fixed sample size ($N = 3000$), for four policy responses per respondent, and for different hypothetical treatment 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

Before respondents saw the issue-position prompt (figure 2), they first saw an introductory screen for the issue at hand. For control group respondents, this introductory screen contained only a short description of the issue at hand (the blue text visible in figure 1), while for treatment group respondents the introductory screen contained both the description of the issue as well as the open-ended reason-giving prompt depicted in figure 1. In this section, I analyse the amount of time that respondents in each group spent on this introductory screen as measure of engagement with the issue at hand before respondents provided their responses to the issue-position questions. Note that duration data was only collected for the first wave of the survey, and so the results in this section are presented only for responses collected during that wave.

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 difugre demonstrates that in the first wave of the survey, 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.

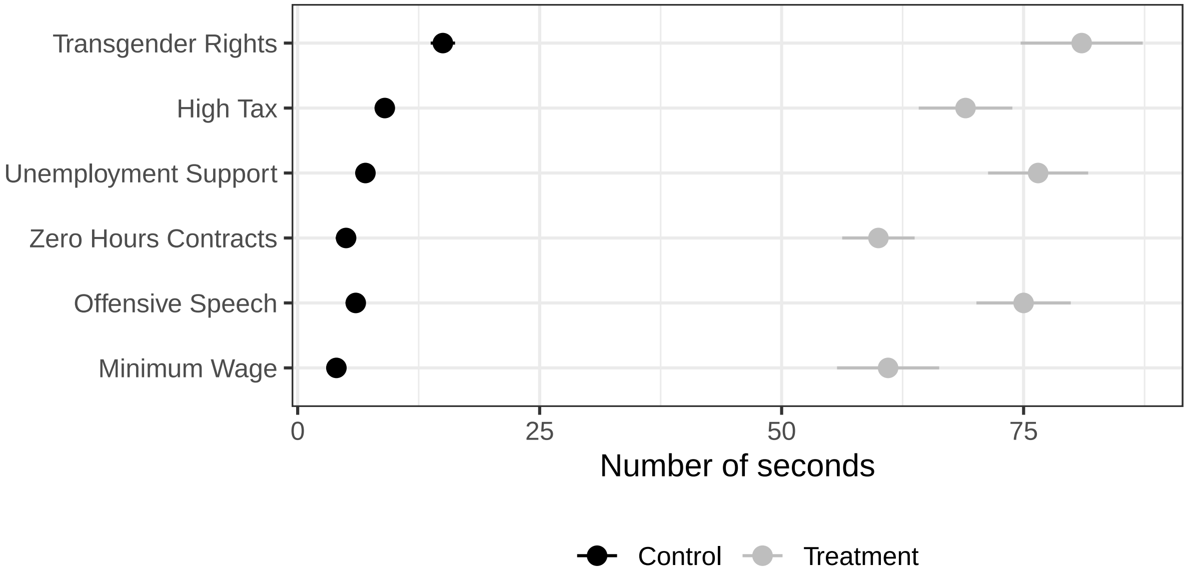


Figure A2: Median introductory screen duration per issue for treatment and control groups

Figure A3 plots the distribution of the number of seconds that treatment group respondents spent on the introductory screen for each issue, in bins of fifteen seconds. The plot demonstrates that, while there is a large degree of heterogeneity in the amount of time that treatment group respondents engaged with the reason-giving task, the vast majority of treatment group units spent more than 15 seconds on the introductory screen. Given that the median duration for control units on the introductory screen was between 4 and 15 seconds, this implies that between 93% and 99% of treatment group respondents spent more time thinking about the issue at hand than did the typical control group respondent, depending on the issue. Across all issues, this distribution is positively skewed, reflecting the fact that a small number of respondents spent a very long time on the introductory screen.

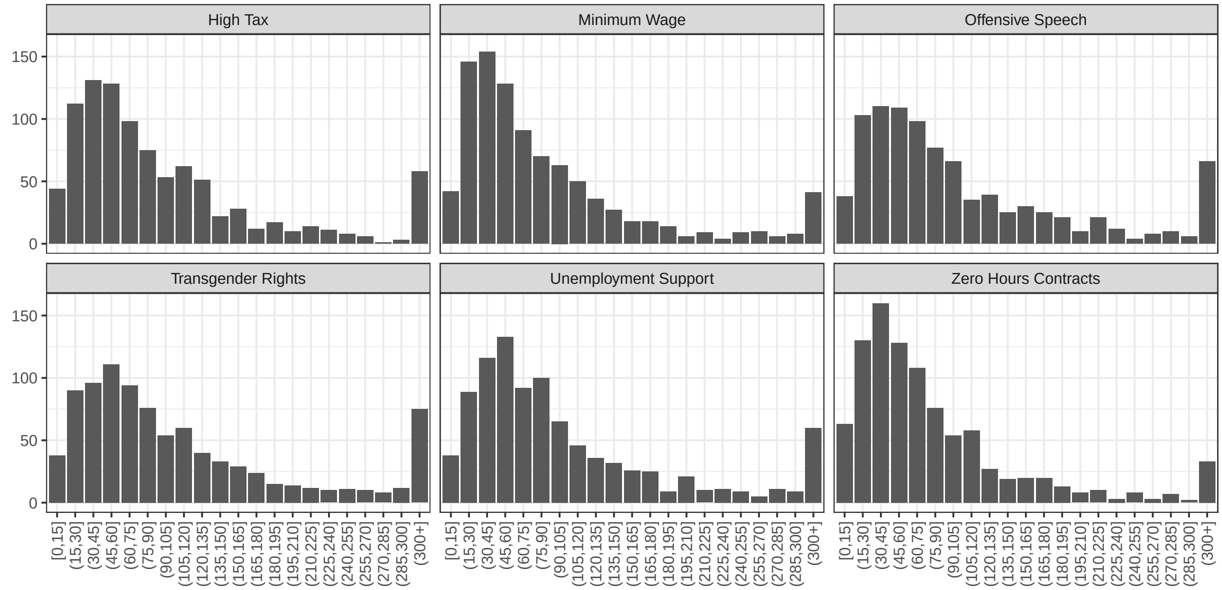


Figure A3: Introductory screen duration per issue, binned

One potential concern is that differential engagement with the reason-giving task might undermine the conclusions presented in the manuscript. In particular, one might worry that those respondents who spent less time thinking about the reasons for their attitudes might be less likely to shift their attitudes in response to being in the treatment group. While the amount of time that a respondent spends on the introductory screen is not itself randomly assigned, and there are plausible confounders that might jointly determine attentiveness to the reason-giving task and responses to the issue position questions, I nevertheless present results below which condition on this variable. In particular, I subset the treatment group to exclude those responses where the respondent spent less than 30 seconds on the introductory screen for the relevant issue. I then re-estimate the main quantities of interest for the constraint, stability and polarization outcomes and present the results in figure A4.

The figure demonstrates that restricting the treatment group to those respondents who more clearly engaged with the treatment has no substantive effect on the results reported in the paper. The black points and intervals in the figure represent the treatment effect for those who spent longer than 30 seconds on the introductory screen, and the grey points represent the treatment effects for the full sample as reported in the main body of the paper. The estimated treatment effects are substantively very similar and statistically indistinguishable.

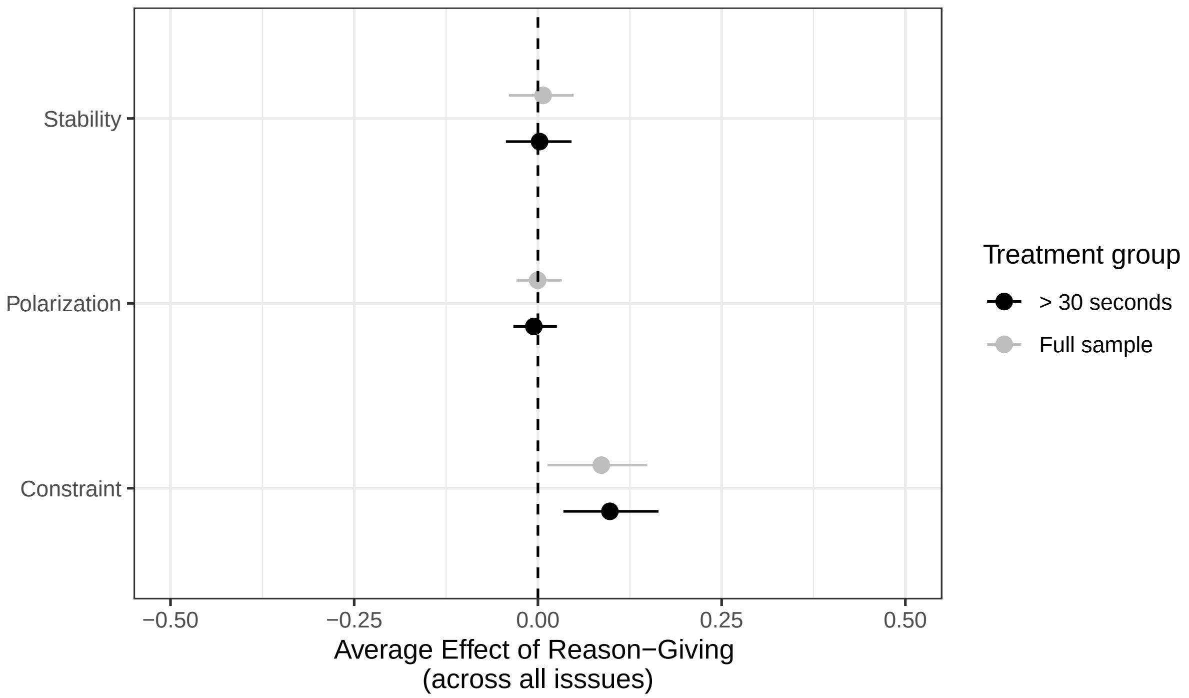


Figure A4: Average effect of reason-giving for treatment units who spent longer than 30 seconds on the reason-giving task

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.

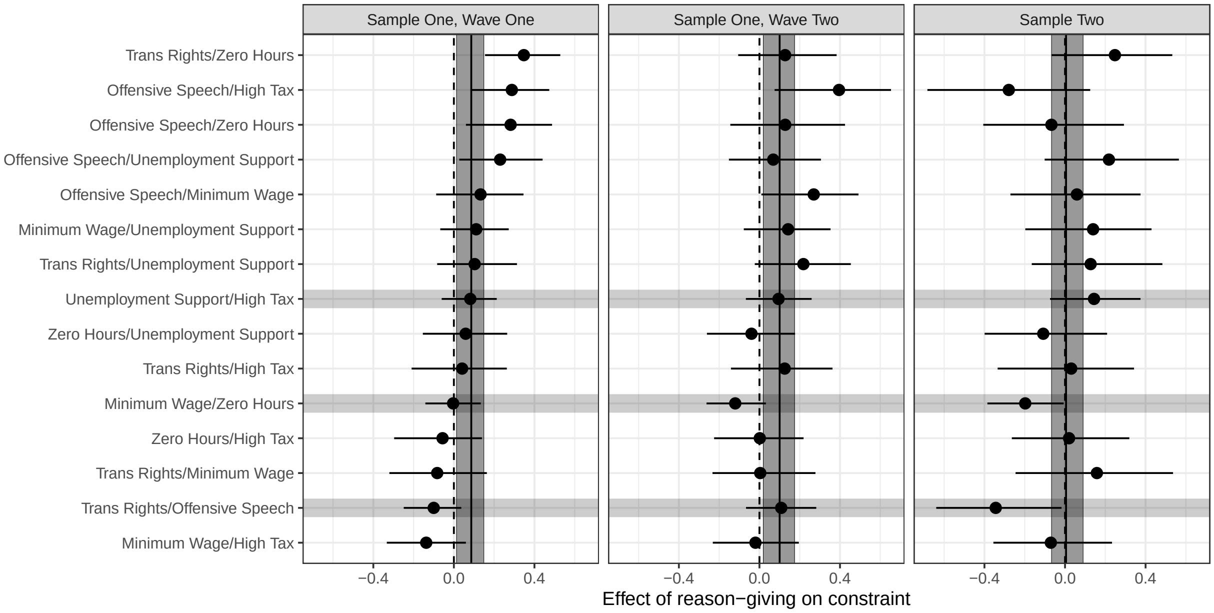


Figure A5: Effects of Reason-Giving on Ideological Constraint (Attrition Weighted)

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 A5, A6, and A7. 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.

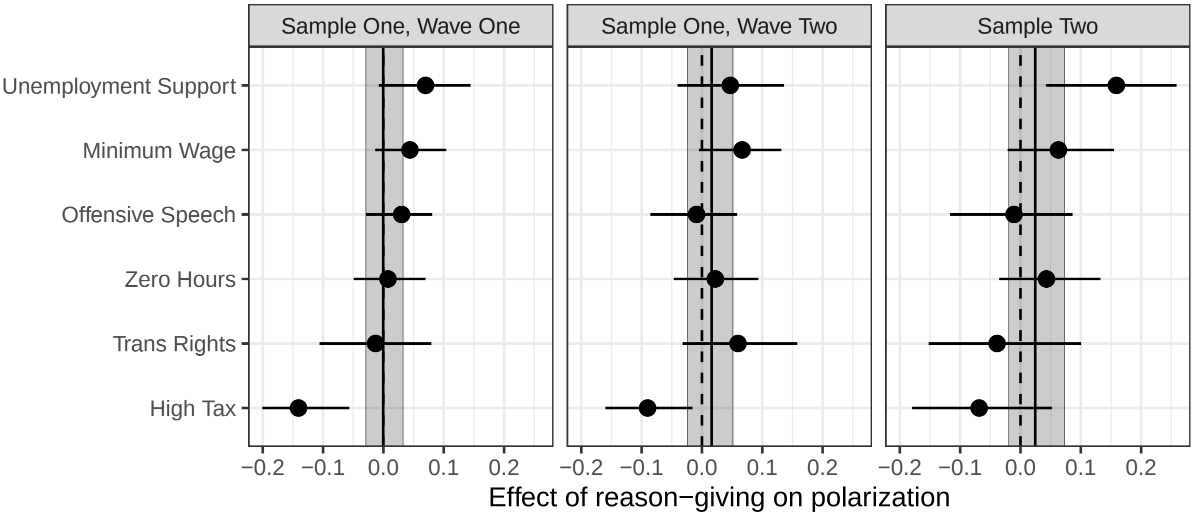


Figure A6: Effects of Reason-Giving on Attitude Polarization (Attrition Weighted)

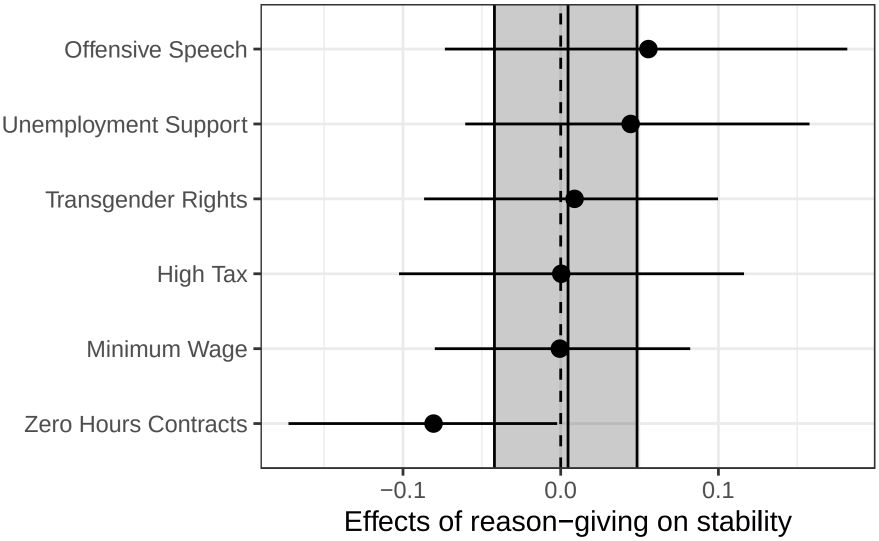


Figure A7: Effects of Reason-Giving on Attitude Stability (Attrition Weighted)

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 A8 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 A9 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 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 A10 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.

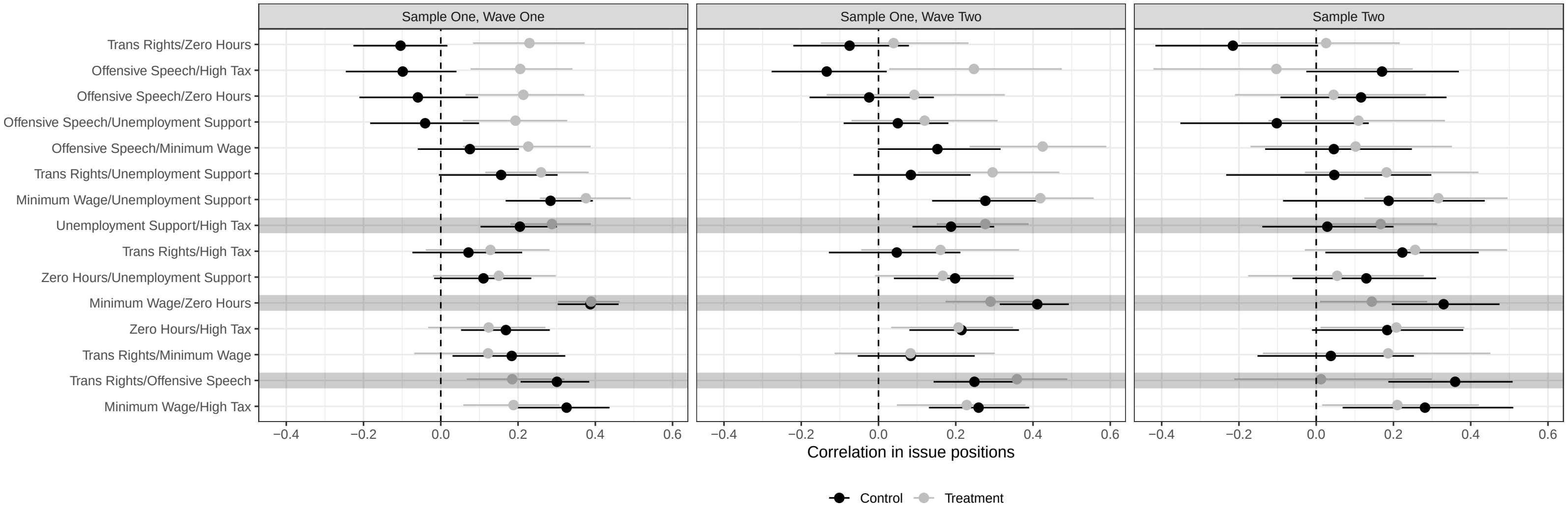


Figure A8: Treatment- and control-group issue-pair correlations

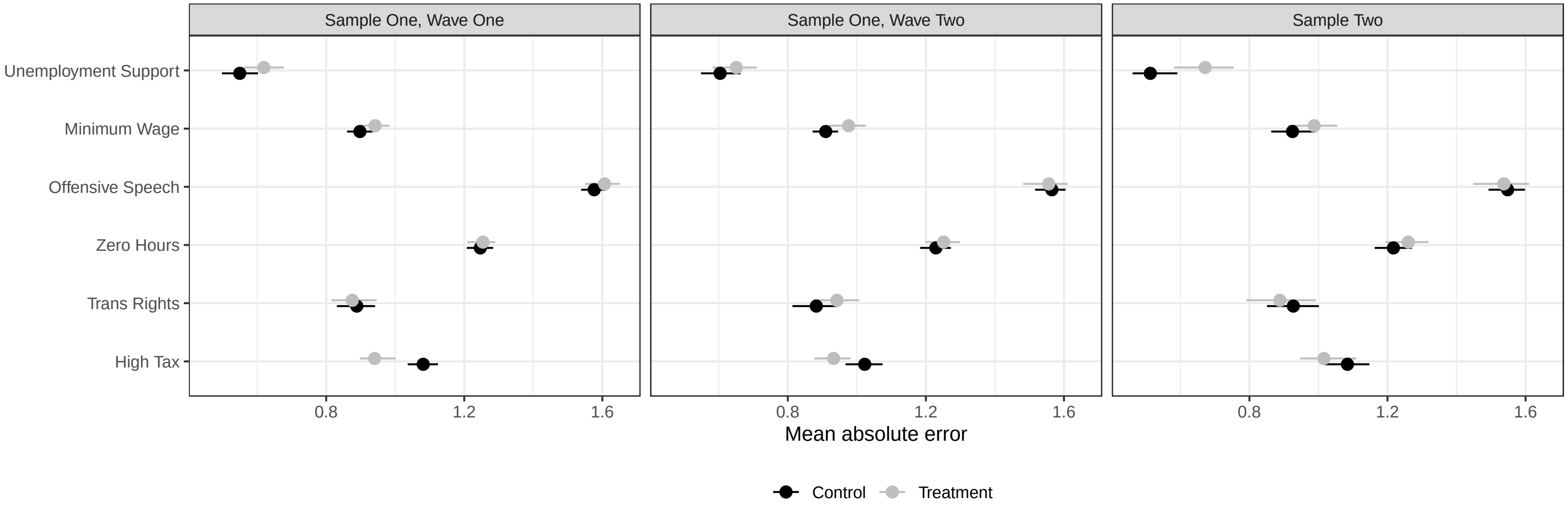


Figure A9: Mean absolute error (treatment and control)

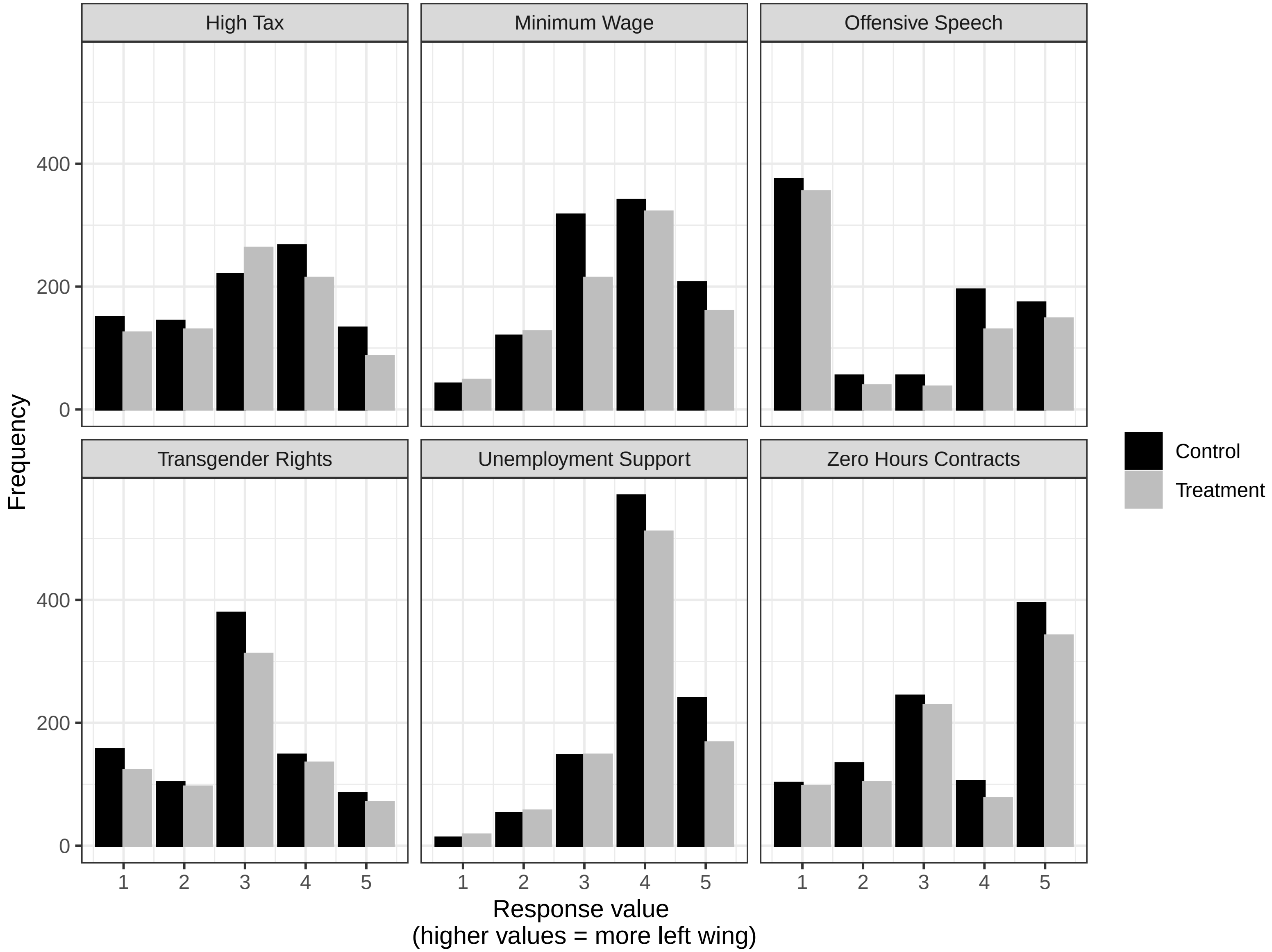


Figure A10: Raw outcome distributions (Sample One, Wave One)

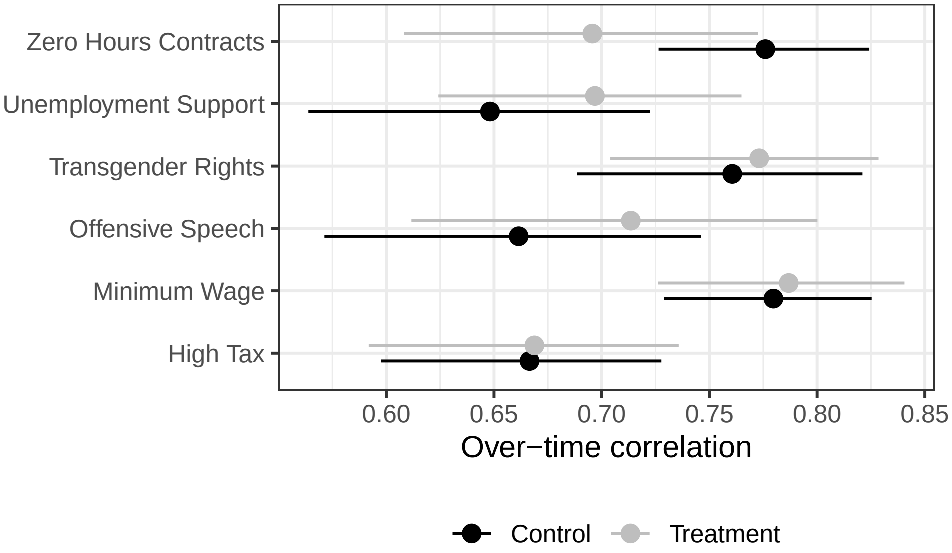


Figure A11: Treatment- and control-group over-time correlations

Stability: Figure A11 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.

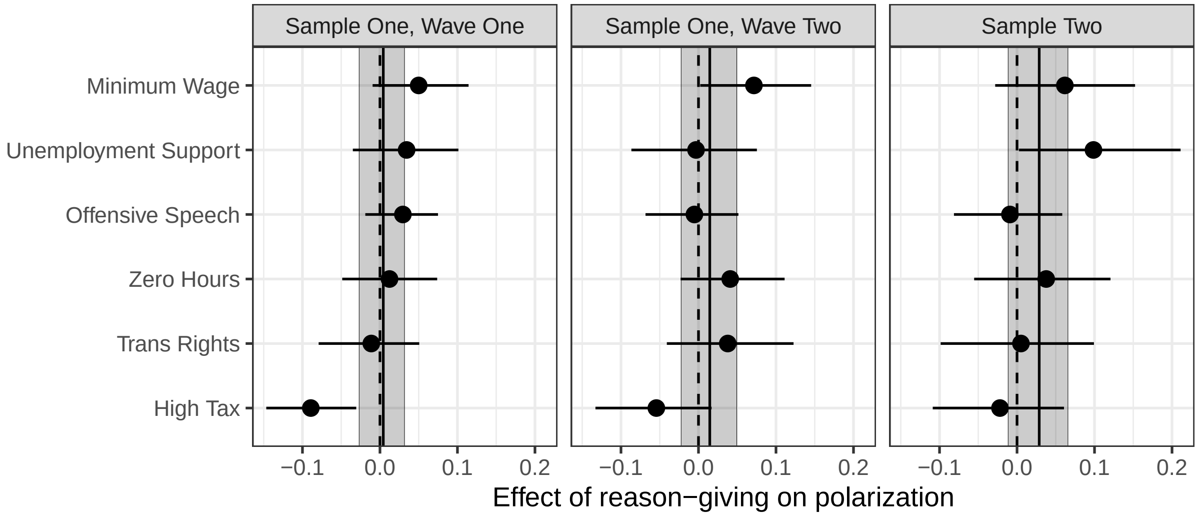


Figure A12: Effects of Reason-Giving on Polarization (Standard Deviation)

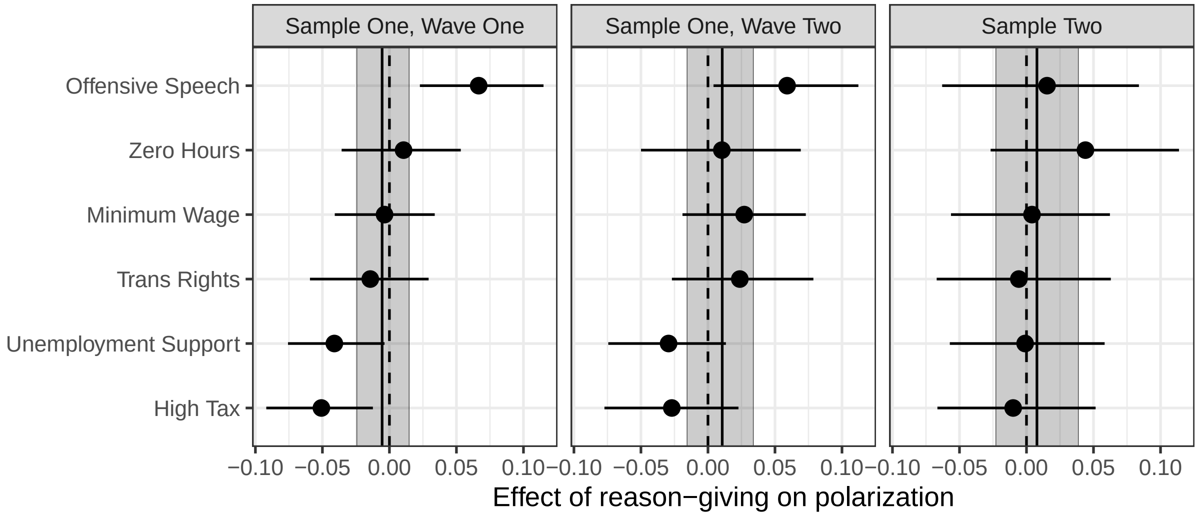


Figure A13: Effects of Reason-Giving on Polarization ("Extreme" responses)

Using these measures I then rerun the analyses depicted in figure 5 of the main body of the paper. Figure A12 depicts the estimated treatment effects using the standard deviation measure, and figure

A13 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 A14 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.

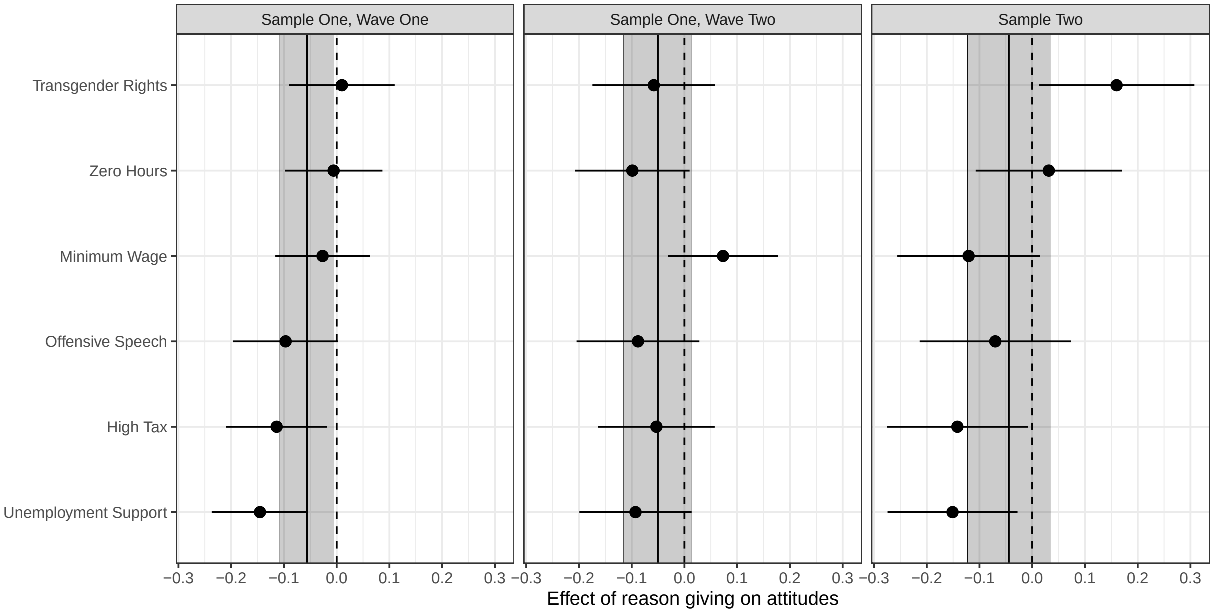


Figure A14: Effects of reason-giving on left-right position

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 A15 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.[13] 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.

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

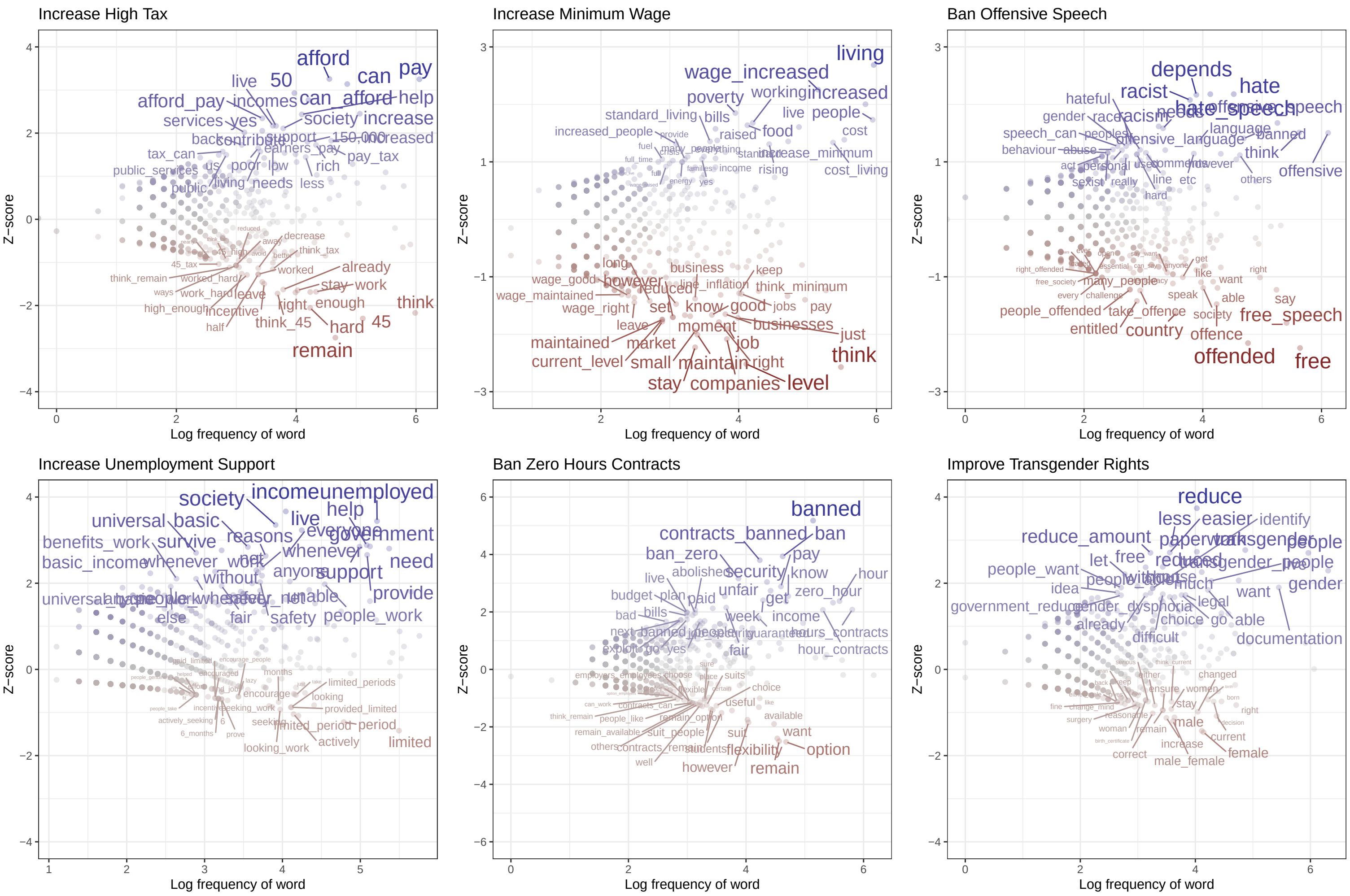


Figure A15: Distinctive token use by issue position

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.

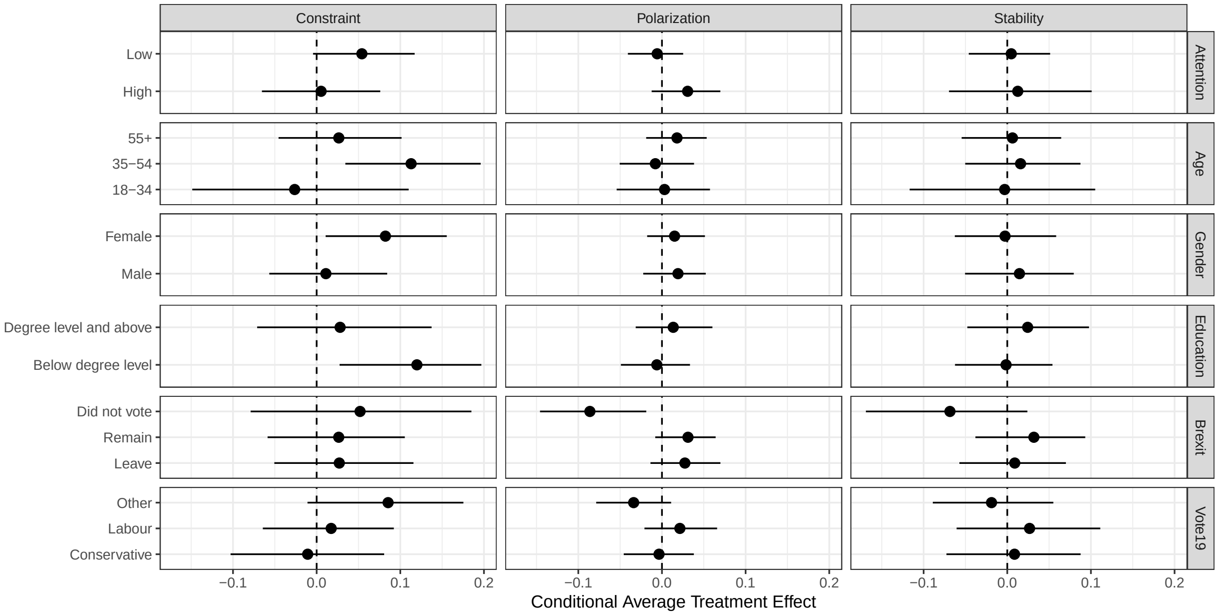


Figure A16: Conditional Average Treatment Effects by Respondent Characteristics

I Heterogeneous Treatment Effects by Voter Characteristics

In analyses that were not pre-registered, figure A16 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 little evidence of 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.