

## Appendix Table of Contents

### Contents

---

<b>A</b>	<b>Survey Prompts</b>	<b>A2</b>
<b>B</b>	<b>Power Analyses</b>	<b>A8</b>
<b>C</b>	<b>Question Duration by Treatment Group</b>	<b>A10</b>
<b>D</b>	<b>Item and Unit Non-Response</b>	<b>A12</b>
<b>E</b>	<b>Ceiling and Floor Effects</b>	<b>A15</b>
<b>F</b>	<b>Alternative Measures of Polarization</b>	<b>A20</b>
<b>G</b>	<b>Treatment Effects on Left-Right Preferences</b>	<b>A22</b>
<b>H</b>	<b>Reasons Given</b>	<b>A24</b>
<b>I</b>	<b>Heterogeneous Treatment Effects by Voter Characteristics</b>	<b>A26</b>

---

## 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

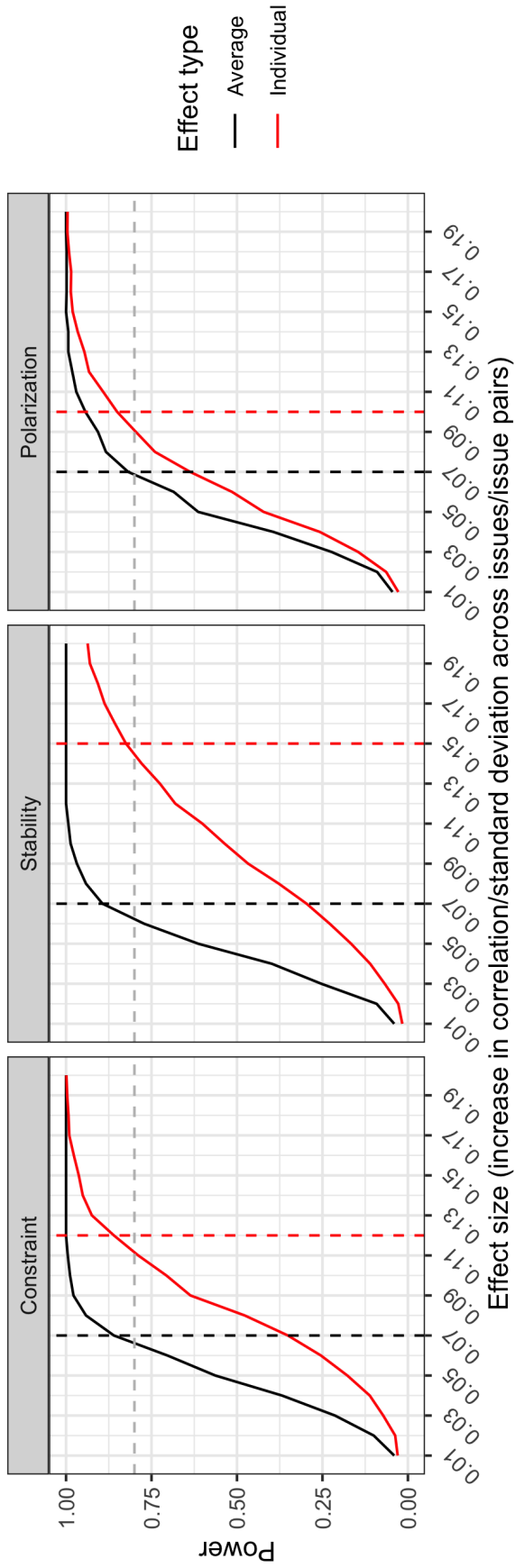
Figure A1 shows the results of a power analysis for the quantities of interest described in section 4 of the paper. To construct the power analysis, I simulated the data collection process for a fixed sample size ( $N = 3000$ ), for four policy responses per respondent, and for different hypothetical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

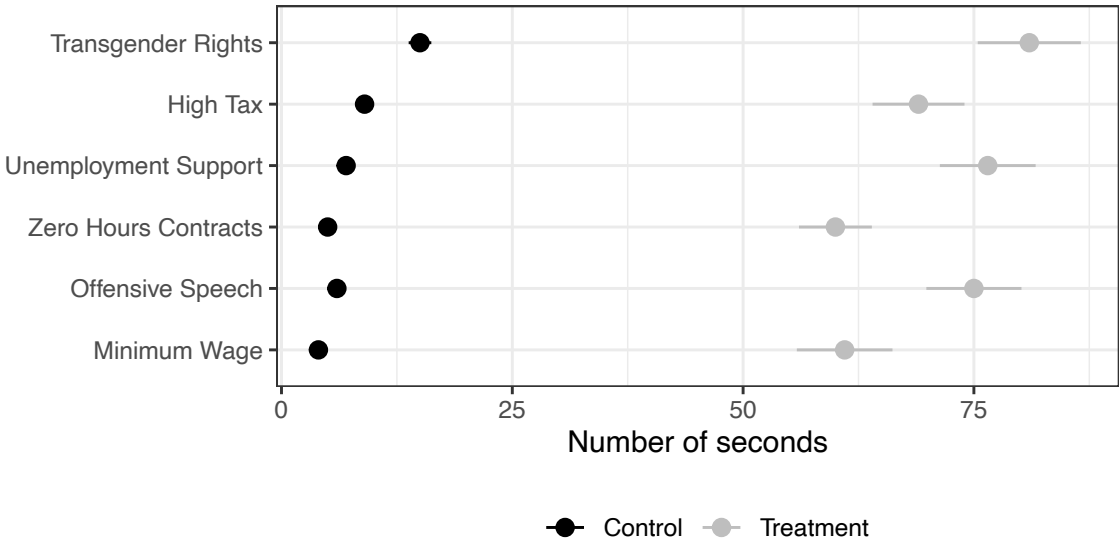


Figure A2: Median introductory screen duration per issue for treatment and control groups

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.

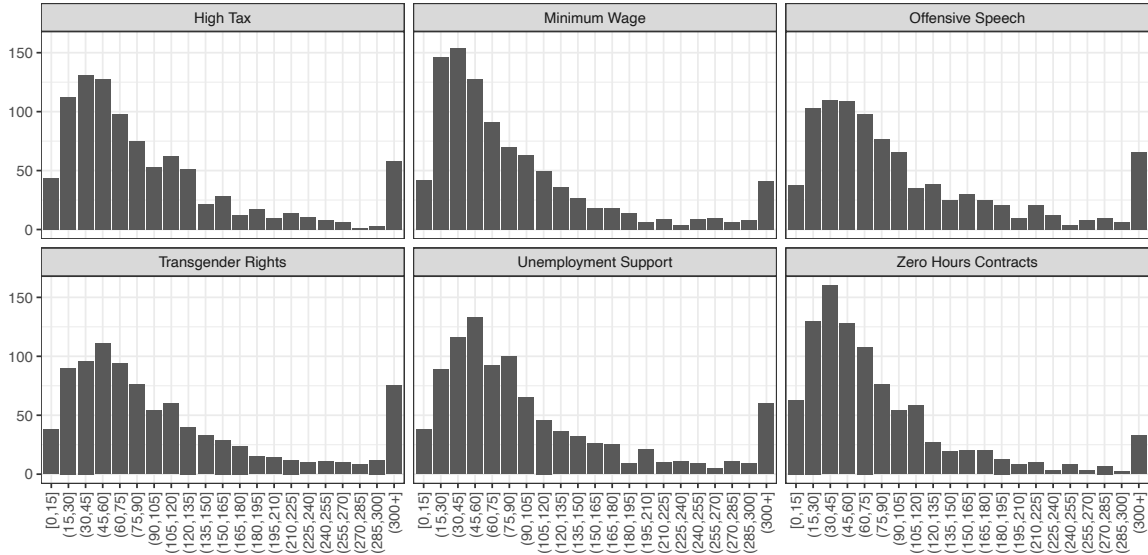


Figure A3: Introductory screen duration per issue, binned

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.

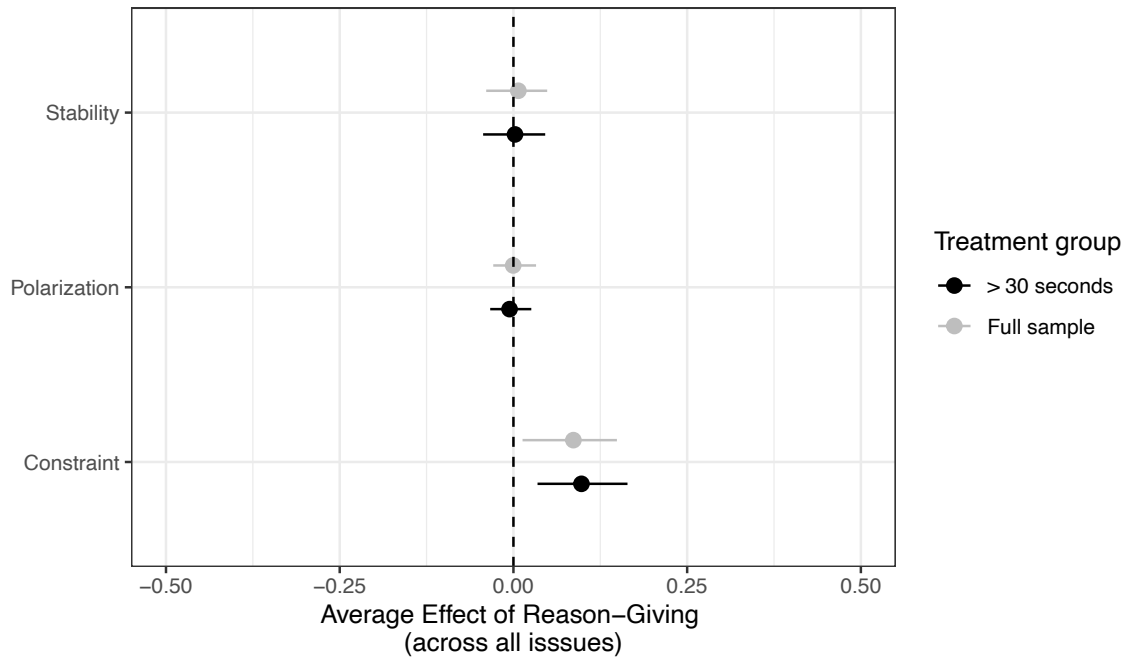


Figure A4: Average effect of reason-giving for treatment units who spent longer than 30 seconds on the reason-giving task

## 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.

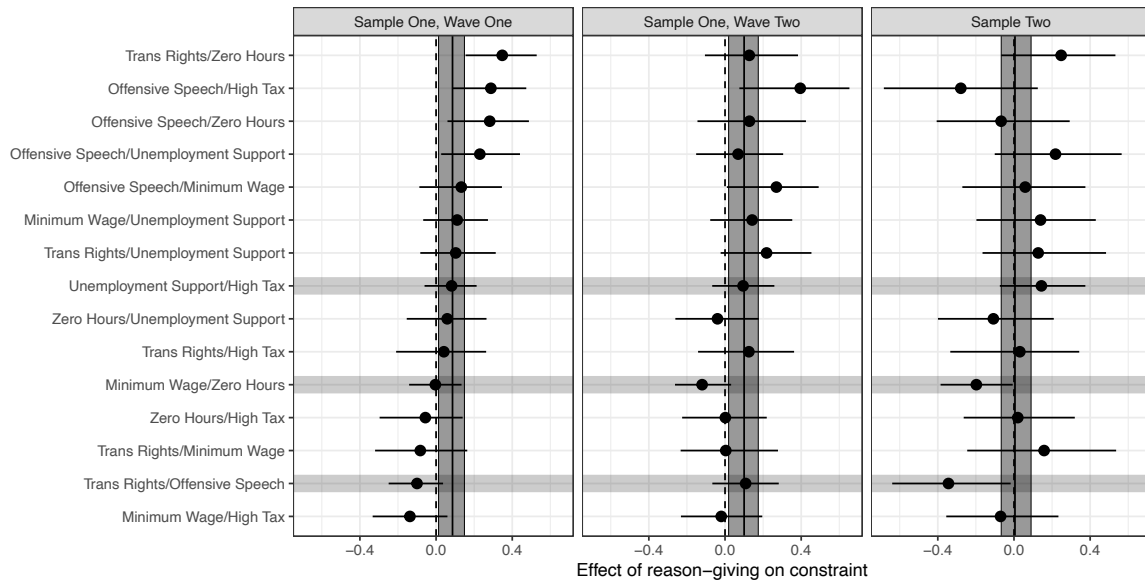


Figure A5: Effects of Reason-Giving on Ideological Constraint (Attrition Weighted)

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.

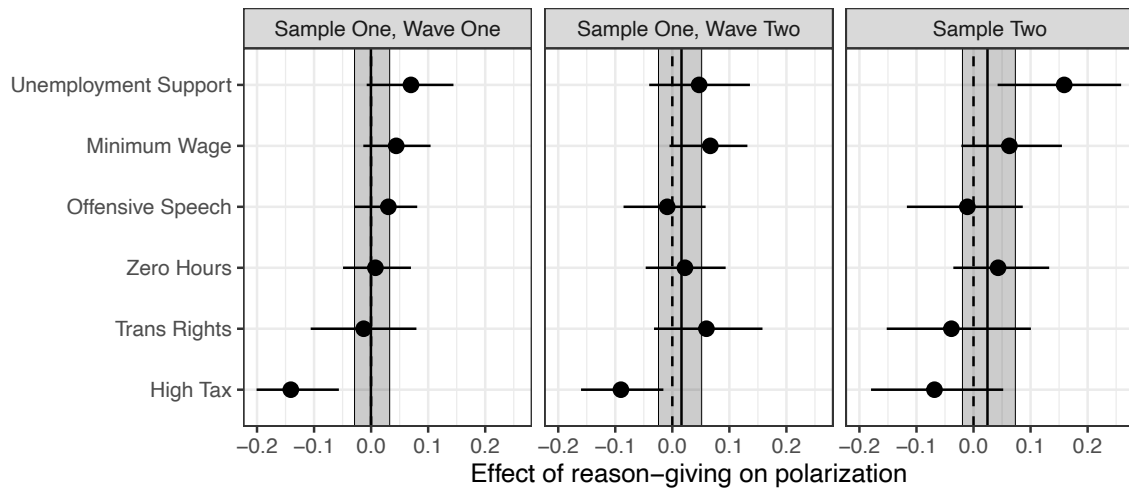


Figure A6: Effects of Reason-Giving on Attitude Polarization (Attrition Weighted)

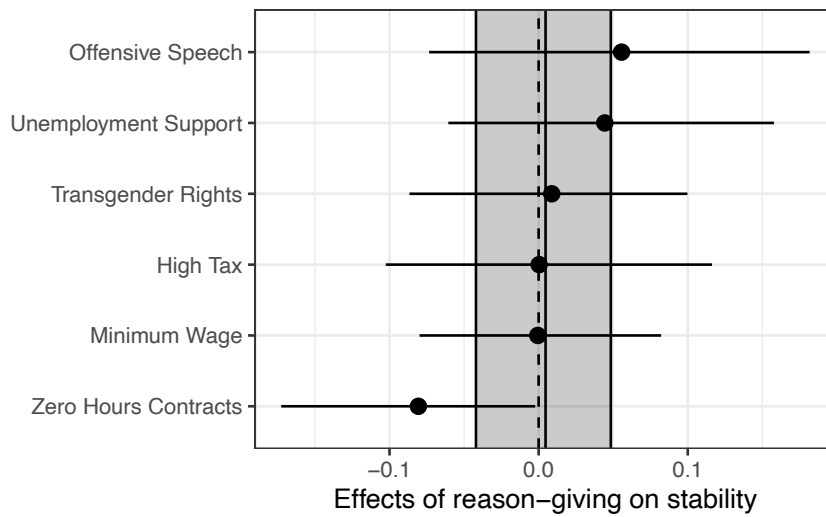


Figure A7: Effects of Reason-Giving on Attitude Stability (Attrition Weighted)

## 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.

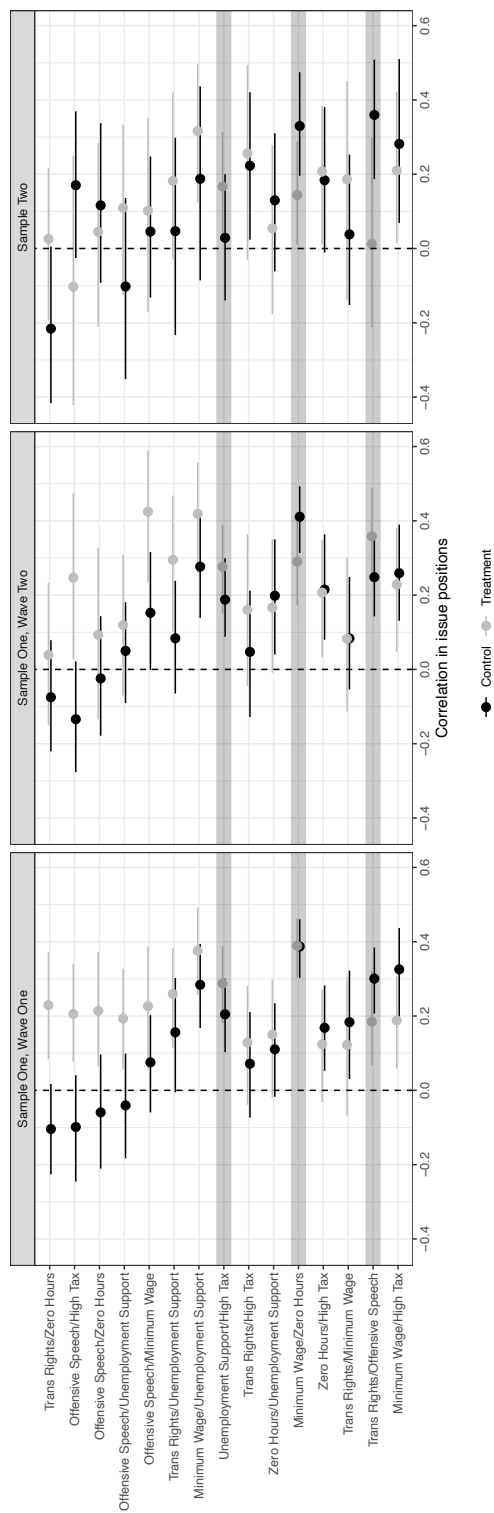


Figure A8: Treatment- and control-group issue-pair correlations

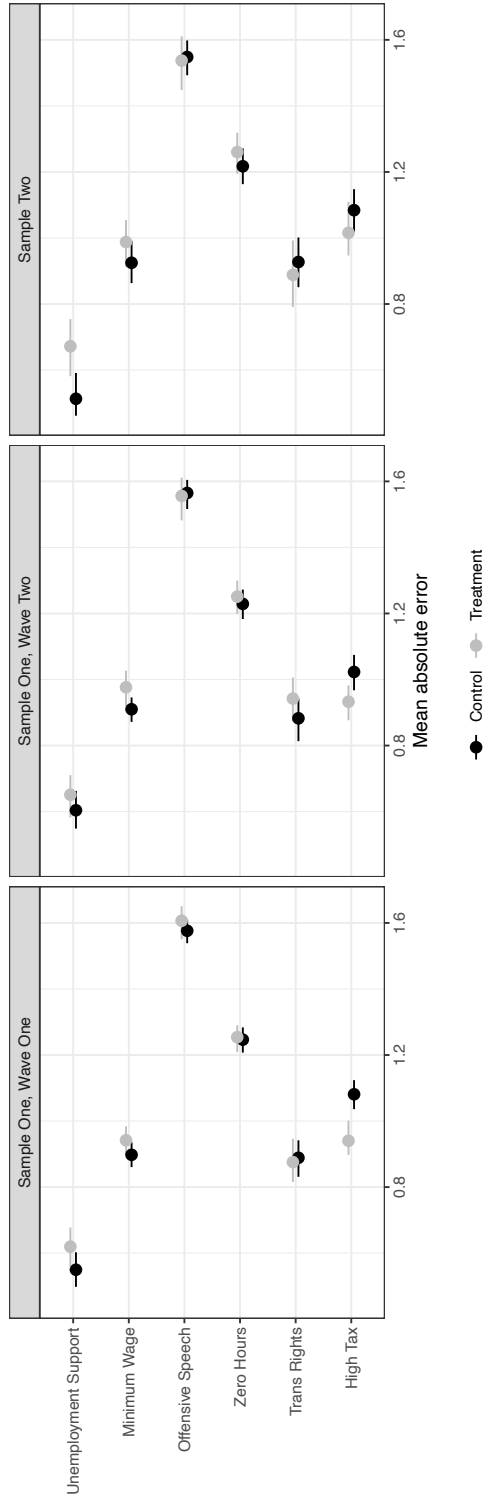


Figure A9: Mean absolute error (treatment and control)

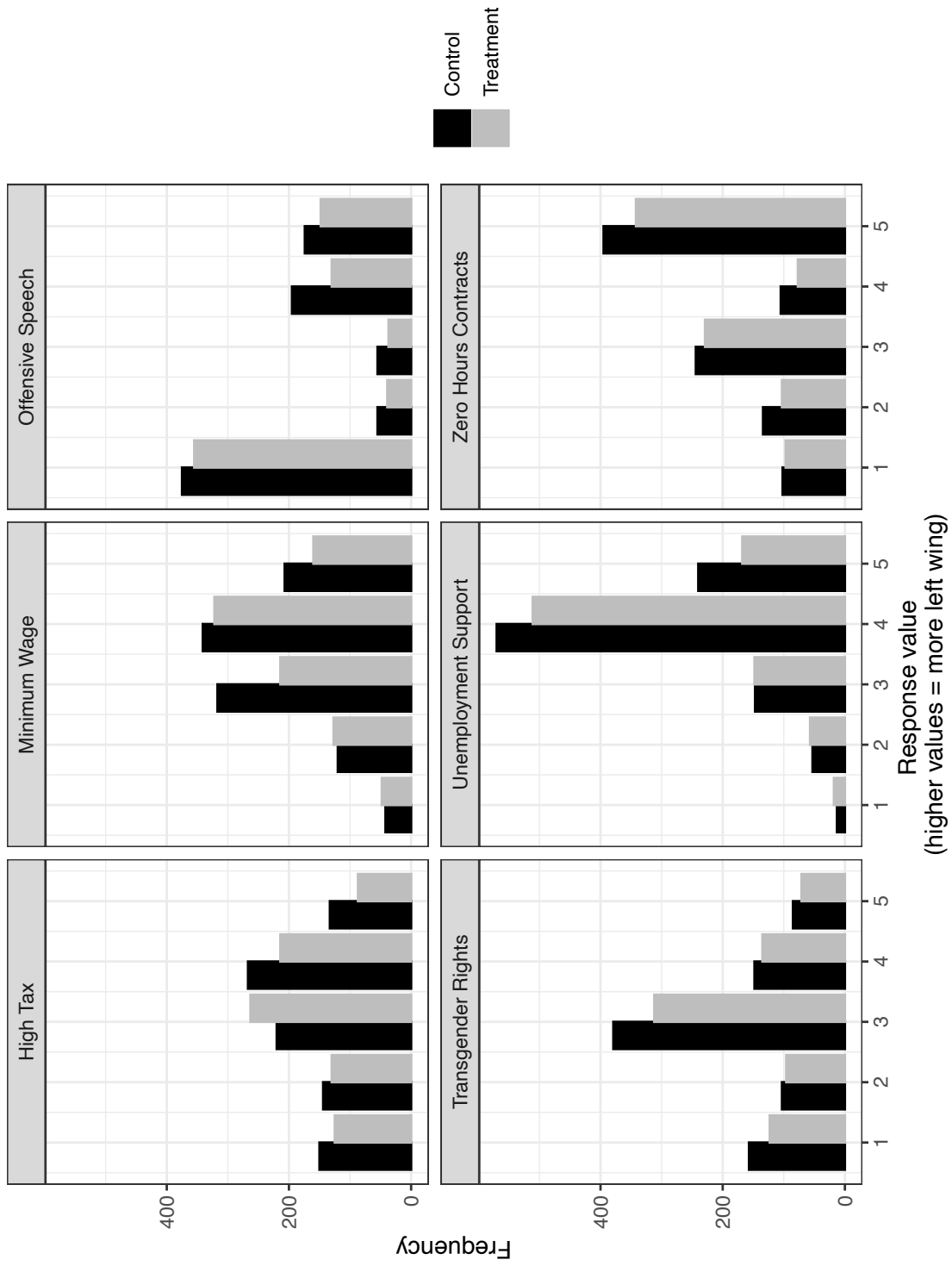


Figure A10: Raw outcome distributions (Sample One, Wave One)  
 (higher values = more left wing)

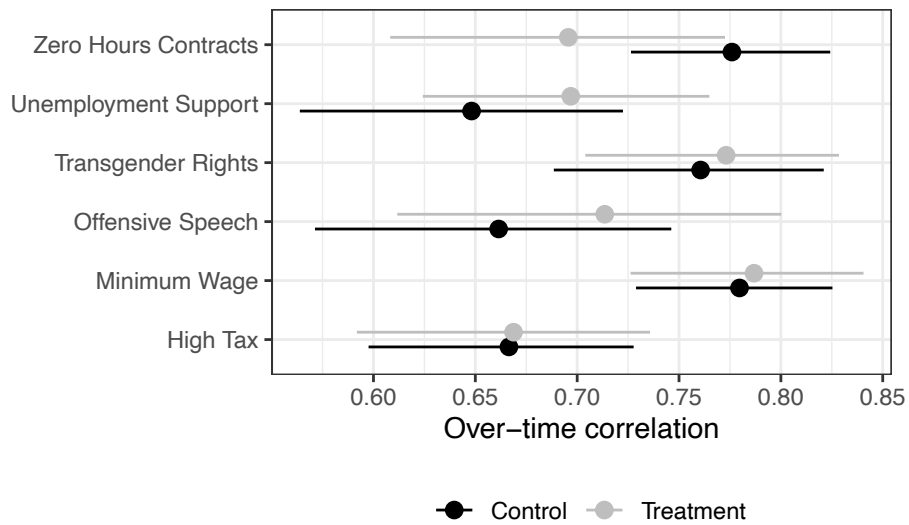


Figure A11: Treatment- and control-group over-time correlations

*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.

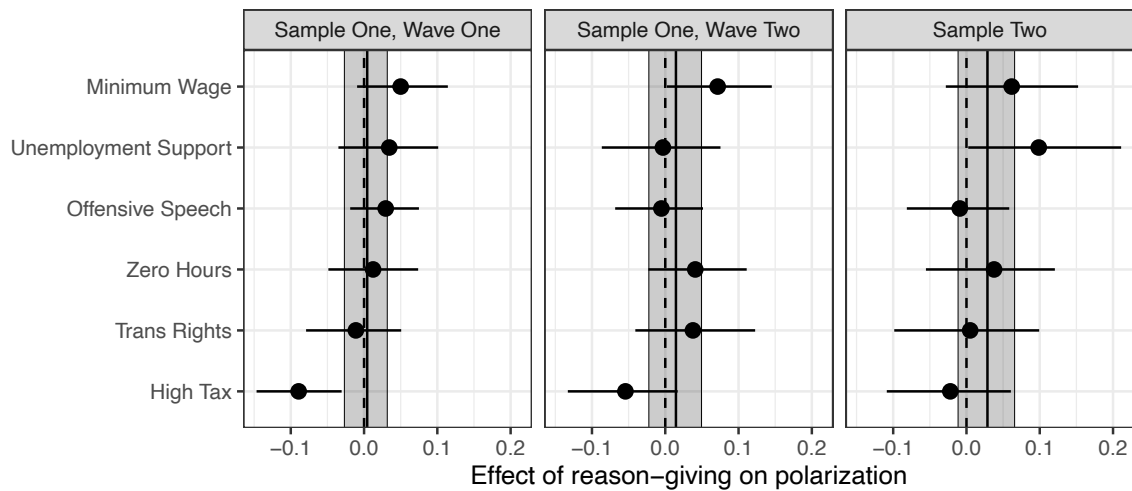


Figure A12: Effects of Reason-Giving on Polarization (Standard Deviation)

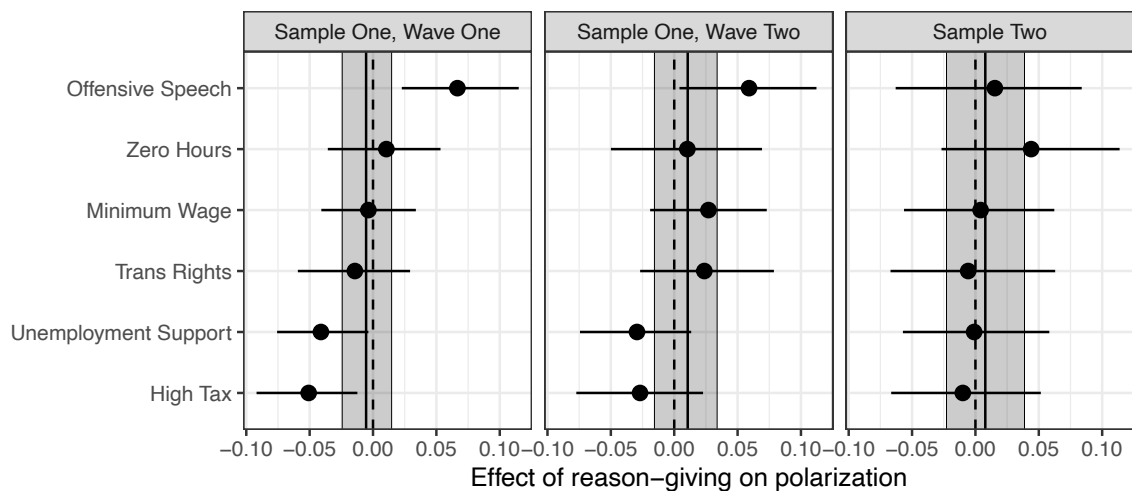


Figure A13: Effects of Reason-Giving on Polarization ("Extreme" responses)

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.

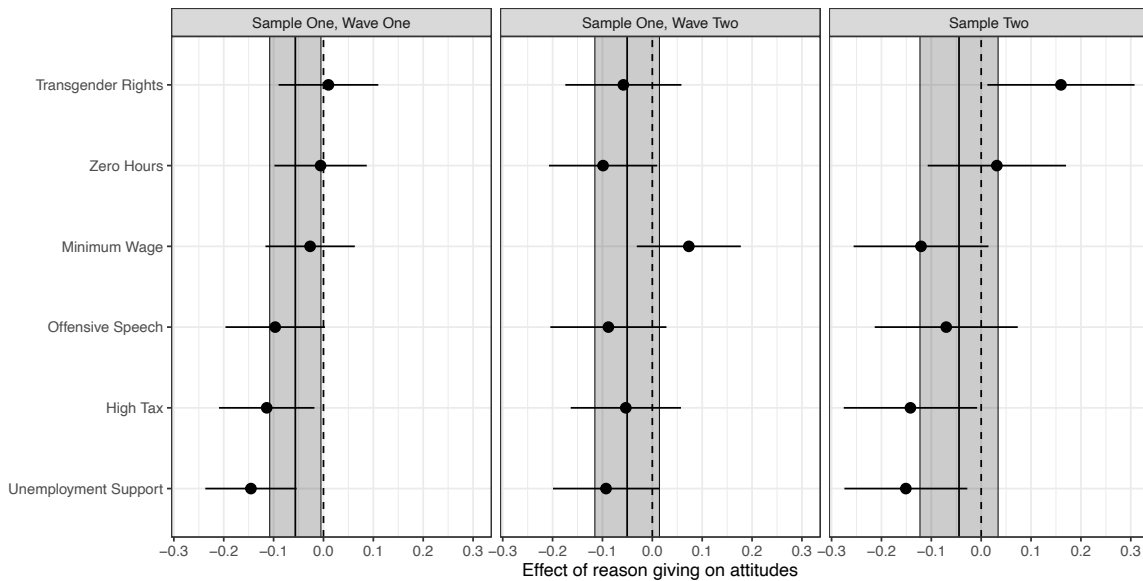


Figure A14: Effects of reason-giving on left-right position

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.<sup>13</sup> 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.

---

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



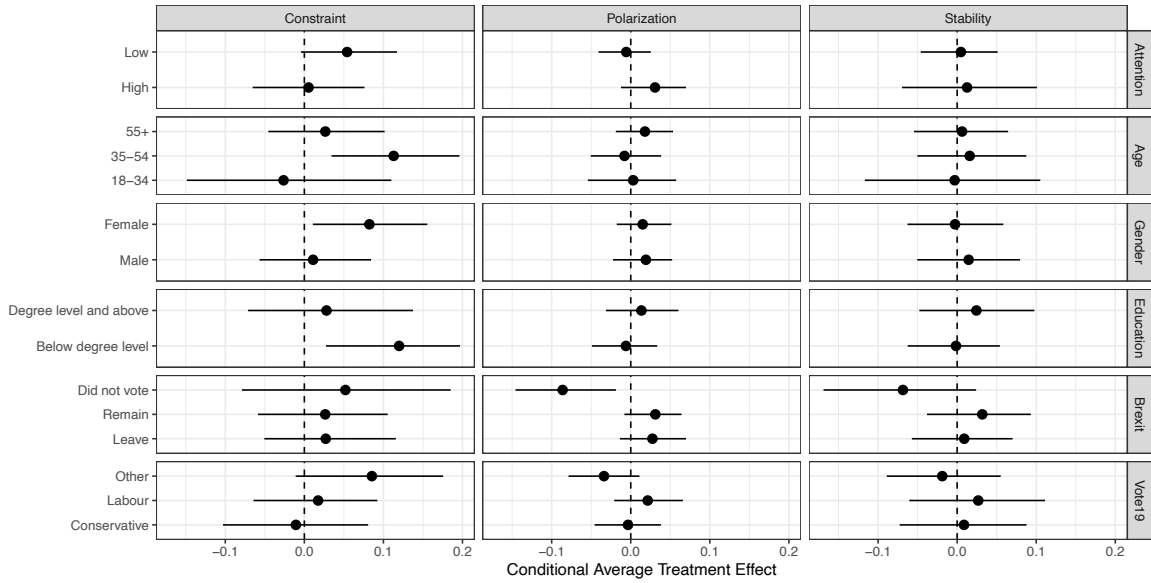


Figure A16: Conditional Average Treatment Effects by Respondent Characteristics

## 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.