# Supplemental Materials: A Text-As-Data Approach for Using Open-Ended Responses as Manipulation Checks

Jeffrey Ziegler[†]

[†]Institute for Quantitative Theory and Methods, Emory University, Atlanta, GA 30322, United States. E-mail: jeffrey.ziegler@emory.edu.

The first portion of the Supplemental Materials (Section SM.1) presents the benefits and drawbacks of open-ended responses in comparison to closed-ended responses. I then discuss the basic intuition behind document similarity measures and how they are calculated in Section SM.2. I show that the similarity measures used in the manuscript are highly correlated with other commonly used measures of text similarity, including word embeddings, as well as with factual correctness. Third, I describe the benefits of using weights to diagnose the impact of attention on the overall treatment effect, i.e. PATE (Section SM.3). I also discuss how I simulate the LATE for those participants that likely received the treatment in Section SM.4. I include supplementary information for the re-anysis of the survey experiment that I conduct in the manuscript in Section SM.5. Last, I show how to implement open-ended manipulation checks in R using the package I developed with an additional application conducted in Brazil and Mexico (Section SM.6).

## SM.1 Pros and Cons of Open-Ended Responses

Though open-ended responses have been shown to tap into the same underlying attitudes as close-ended items (Geer 1991; Krosnick 1999), close-ended questions are still more popular largely because they are cheaper and easier to code (Presser and Schuman 1996). This applies as well to the application of manipulation checks in which it has

been relatively rare for researchers to use open-ended manipulation checks instead of instructional or factual closed-ended manipulation checks. In the absence of general use, however, social scientists have still constructed clearer measures of participants' open-ended responses to manipulation checks.

For example, Banks and Valentino (2012), Friedman and Sutton (2013), and Clifford and Jerit (2014) all use open-ended manipulation checks; and Kane and Barabas 2019 (2019) include open-ended manipulation checks in 50% of their reported experiments. Unfortunately, open-ended responses are often not analyzed in the main text and are relegated to the Appendix given researchers' hesitancy on how to present the results. The central motivation of this paper is to overcome these shortcomings so researchers can maximize the benefits of open-ended responses, specifically to gain insight into how well respondents pay attention to the task at hand. Yet, some issues remain that researchers should consider before using open-ended responses in manipulation checks.

The most prominent criticism of open-ended responses in the context of manipulation checks is that non-responses are due to inability rather than inattention because respondents lack the necessary rhetorical aptitude to answer correctly. This may especially be the case if survey experiments are administered online and respondents must type their responses. It is difficult, however, to determine if the same individuals that are less attentive to an open-ended manipulation check would be "attentive" if we used a close-ended manipulation check because they are truly attentive and lacked ability, not because they can guess more easily.[1] Nevertheless, we can at least check

[1]Ultimately, we cannot compare whether open-ended manipulation checks confuse ability and attention *less* than closed-ended manipulation checks because we cannot know if individuals that appear less attentive would be more attentive if they were presented with a closed-ended manipulation check. Even if we knew how participants would respond to both an open- and closed-ended manipulation check, the

if attention is associated with common demographic characteristics by regressing our measure of attention on socio-demographic variables such as age, race, education (see Section SM.5).[2] In past studies, however, the "few individuals who fail to respond to these questions appear uninterested in politics, and probably would respond if they had reason to" (Geer 1988, 366).

This raises a separate concern that correct responses to open-ended responses may be heavily impacted by interest, not ability (Holland and Christian 2009). If we place open-ended manipulation checks after a treatment that is especially salient, we may violate our assumption that all respondents provide the same level of attention irrespective of the treatment condition that they are assigned to. Importantly, we can check whether this assumption holds empirically.

In Kane (2020), we are specifically concerned that partisans may pay greater attention to prompts that interest them more. For example, Democrats may prefer to read about disunity within the Republican party and thus pay "more attention", while they would be less attentive to a story that they did not want to read. To check, we can regress

lack of variation that closed-ended manipulation checks force with a correct or incorrect answer makes it impossible to establish if someone is (1) attentive and does not have the capability to make it known with an open-ended manipulation check but can make it known with a closed-ended manipulation check, or (2) not attentive and cannot fake being attentive with an open-ended manipulation check but can guess the correct answer with a closed-ended manipulation check.

[2]If we are interested in modeling the latent associated traits of attention (such as age, education), it should actually be easier and more informative when our measure of attention comes from an open-ended rather than closed-ended manipulation check. For example, I briefly checked the percent of respondents that correctly answered factual closed-ended manipulation checks in some recent Political Science publications and found that it was typically above 90%, which does not really distinguish attention between participants though it likely exists (Edwards and Arnon 2019; Keiser and Miller 2020; Jamieson and Weller 2019; Kim and Kweon 2020; Ladam 2019). If there is very little variation in our measure of attention, socio-demographic variables do not have any variation to explain. Therefore, it is difficult to tell with closed-ended manipulation checks whether a true relationship exists between socio-demographic variables and attention, or whether the indicator of attention itself does not capture the full variation that is present.

respondents' attention on the interaction of their treatment assignment and party identification to see if partisans provide different levels of attention by treatment, on average. I show in Section SM.5 that there is not evidence of a relationship between party ID and the treatment, but all researchers should investigate this assumption.

Though these represent some of the limitations of open-ended responses, the benefits of open-ended responses are numerous. First, open-ended responses inherently contain "more exact information than is possible in a closed format. Even with finely graded categories, there is inevitably some loss of information when the answer is categorical" (Tourangeau, Rips and Rasinski 2000, 232). This is especially true if researchers only include one or two closed-ended manipulation checks in which respondents can only be correct or incorrect.

Additionally, respondents can draw inferences about what the correct answer to the manipulation check is based on the answers that are provided. If respondents are presented with more options, they may also then begin to guess and be more likely to select the middle category because participants interpret the middle category as the population average and the end categories as being very rare (Bishop 1987). Given the combined design benefits of open-ended responses and the advantages of similarity measures, which I discuss in the next section, open-ended manipulation checks provide researchers with a viable alternative to closed-ended manipulations.

## SM.2 Similarity Measures in Text

Our goal is to quantify how alike the text that participants read is to the text they provide as part of the open-ended manipulation check. Political scientists have applied *document similarity* measures to uncover commonalities in language to track the origins of policy

proposals in legislation (Jansa, Hansen and Gray 2019; Wilkerson, Smith and Stramp 2015), as well as explore party messaging strategies (Garrett and Jansa 2015). I rely on two approaches to calculate document similarity measures: $n$-grams and word embeddings.

The first step to calculate any $n$-gram document similarity measure is to divide the text into shorter segments, or "grams", because they are computationally efficient for very long text strings, they are easily comparable given their limited range ([0,1]), and they are a metric (Van der Loo 2014, 120).[3] I set $n = 3$ because it is recommended for short text given that the number of $n$-grams encountered in every-day language is usually much less than the possible number of $n$-grams allowed by the alphabet. Each language has its own most common $n_3$ grams, and this process can be adapted to any language that uses a written alphabet. For instance, the case presented below in Section SM.6 includes examples in Spanish and Brazilian-Portuguese.

Prior to creating segments, I pre-process the text, which aims to make the text "less complex in a way that does not aversely affect the interpretability or substantive conclusions of the subsequent model" (Denny and Spirling 2018, 168). This includes removing capitalization and punctuation, but I do not remove common "stop words" since $n$-gram similarity measures rely on all characters in the text. Then, I calculate four common similarity measures and plot their correlation to compare the similarity measures used in the manuscript.

The first of four similarity measures I employ is the Jaccard, which is calculated

[3]Similarity measures can be classified as metric, semi-metric or non-metric. A metric similarity measure must satisfy the following rules: (1) The maximum value is one when two items are identical; (2) When two items differ, the similarity is positive (negative similarities are not allowed); (3) Symmetry: the similarity of objects $A$ to object $B$ is the same as the similarity of $B$ to $A$; and (4) Triangle inequality axiom: With three objects, the similarity between two of these objects cannot be larger than the sum of the two other similarity (McCune, Grace and Urban 2002, 46).

as the size of the intersection divided by the size of union of two sets. For example, consider the two statements "make love not war" and "make war not love", which consist of the same words, but they have a Jaccard similarity of approximately 0.58 (there are 11 common grams, divided by the total number of grams, 19). Second, I consider the cosine of the angle, which does not discount similarity based on length. To make this work, all documents, including open-responses and prompts, are stored as sparse vectors (i.e. they have many zeroes) and the overlapping angle between that respondent's written recall and the text that the respondent viewed as the treatment is the cosine similarity.

The next *n*-gram similarity measure I use is the Jaro, which was originally developed by the U.S. Bureau of the Census to link records based on inaccurate text fields. The Jaro similarity should uncover character discrepancies that are caused by typing-errors, so matches between characters further from each other on the keyboard are unlikely to be caused by a typing error. The similarity, therefore, measures the number of matching characters between two strings that are not many positions apart, and adds a penalty for matching characters that are transposed. The last measure I include is the Damerau-Levenshtein, which calculates the similarity between two words as the minimum number of insertions, deletions, or substitutions of a single character, or the transposition of two adjacent characters that are required to change the first word into the second.

Though an *n*-gram representation of words allows for fast computation and comparison, it does not capture the meaning of individual words or sentences. For example, take the sentences "Obama speaks to the media in Illinois" and "The President greets the press in Chicago". While these two statements have no words in common, they

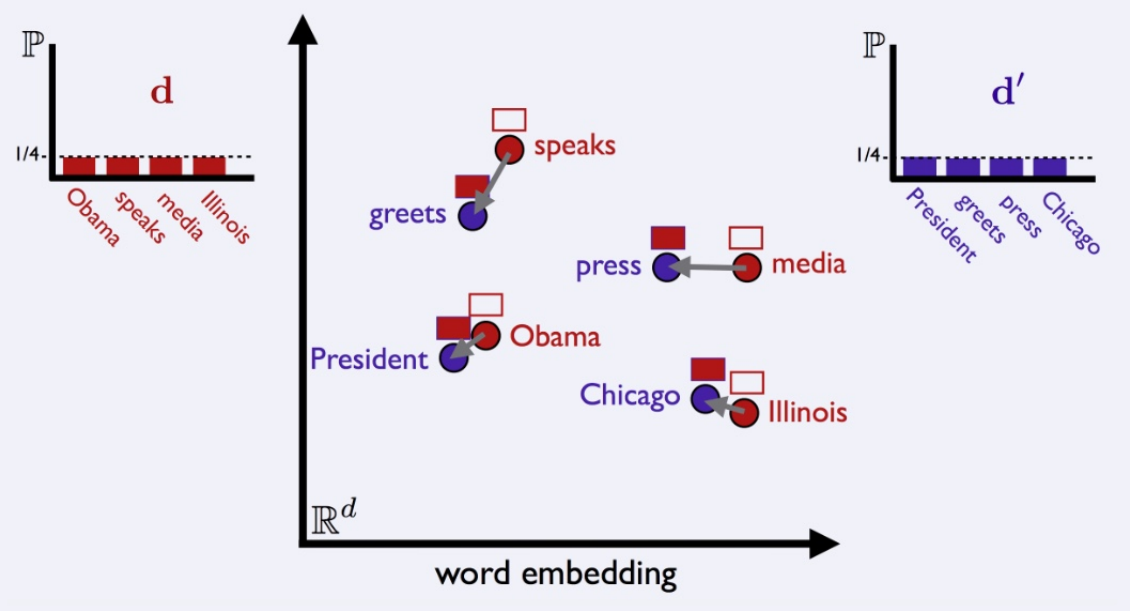
convey very similar information. In this case, the proximity of the word pairs: (Obama, President); (speaks, greets); (media, press); and (Illinois, Chicago) is not accounted for in the $n$-gram similarity measures. To overcome this potential shortcoming of $n$-gram similarity measures, I use Word Mover's Distance (WMD) which relies on trained data to estimate semantically meaningful representations for words from co-occurrences in sentences (Kusner et al. 2015).

For instance, Figure SM.1 uses the example from above to show that distances between words in the embedding space are semantically meaningful. This process works by treating both documents as a weighted point cloud of embedded words. The distance between two texts is calculated by the minimum cumulative distance that words from document 1 need to travel to match exactly the point cloud of document 2. In other words, the WMD algorithm calculates the most efficient way to "move" the distribution of words from document 1 to the distribution of words in document 2.

Figure SM.2 displays the bivariate correlations between all of the similarity measures, including those created by the word embeddings.[4] The correlation between the cosine of the angle using the $n$-gram approach and the cosine of the angle of the word embeddings is 0.87. Importantly, the correlation between the "correct" answer and the cosine of the angle of the word embeddings ($r$ = 0.68) is comparable to the two $n$-gram measures used in the manuscript ($r$ = 0.68, 0.74). Therefore, given the speed and ease of calculating $n$-gram measures, I use them instead of the word embeddings in the manuscript.

[4]Though the Jaccard similarity only takes a unique sets of grams for each response, the cosine of the angle between two vectors considers the total length of the vectors and it can, therefore, be used with the $n$-gram approach or word embedding method. Word Mover's Distance, however, uses a Euclidean distance, which requires a normalization so that the word embedding measure can be compared to the $n$-gram measure.

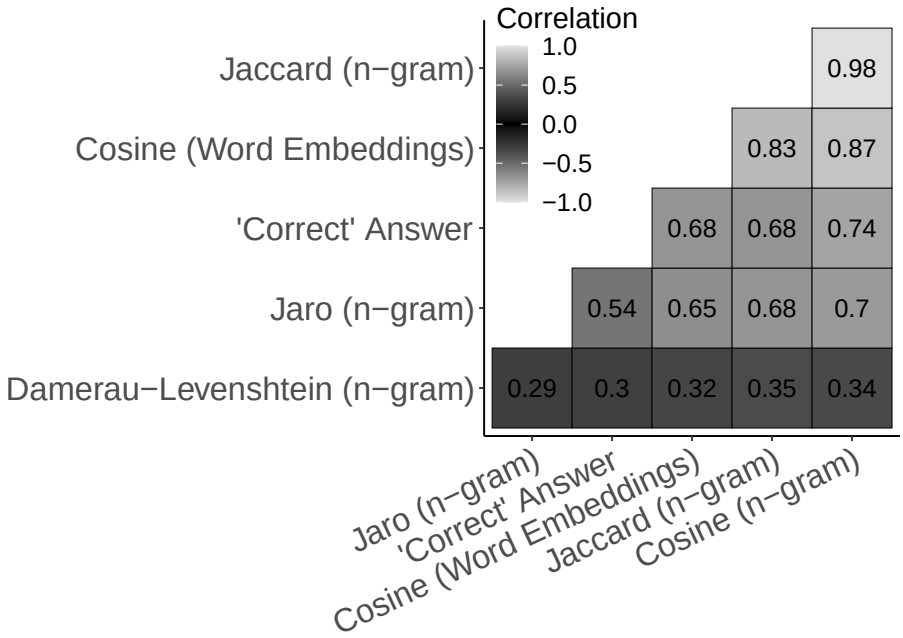
**Figure SM.1:** Comparison of example sentences using Word Mover's Distance.




*Notes:* This example and figure comes from Niculae and Kushner 2015. The meaningful words in the two sentences are shown next to their synonyms, which signals that the cumulative distance between the sentences is low and semantic proximity is high.

Finally, similarity measures of open-ended responses to manipulation checks can be used for other mediums aside from text via online survey experiments, such as in-person interviews, telephone surveys, or experiments that utilize audio. Although audio treatments are not as frequently used as text alone, there are numerous Political S™cience articles that employ an audio component in their treatment (Brierley, Kramon and Ofosu 2020; Hopkins 2015; Iyengar et al. 2008; McClendon and Riedl 2015; Weber and Thornton 2012 to name a few). Given that audio data contains textual information, open-ended manipulation checks could be especially useful. The first step is to convert the audio prompt/treatment as well as participants' open-ended responses in the form of textual transcriptions and audio files. Moreover, there is a growing literature regarding

**Figure SM.2:** Correlation between similarity measures from Kane (2020).




*Notes:* All correlation coefficients are statistically differentiable from zero ($p < 0.05$).

the methodological techniques to assess audio (Dietrich, Mondak and Williams 2020; Knox and Lucas 2020), so it may also possible for researchers to calculate similarity measures of acoustic patterns in auditory open-ended responses, not only textual similarity (Foote 1997).

## SM.3 Weighted Regression Using Similarity Measures

One of the central assumptions of linear regression is that all errors have the same probability density function and the same variance. This assumption is unlikely to be met when all respondents have varying levels of attention. This is problematic because it is

more difficult to obtain unbiased estimates of the overall average treatment effect among the general population (PATE), which means that the PATE will differ from the LATE, or the treatment effect among those individuals that actually "received" the treatment. To address this, we want to account for the probability of receiving the observed treatment independent of the observed covariates, which is precisely what our attention measure captures: those who are less attentive are less likely to have received the treatment and we may expect that they do not represent the average individual that pays greater attention.

As such, we can use weighted linear regression, which we typically rely upon when we want to calculate the correct parameter estimates under endogenous sampling.[5] This exact process occurs when the errors are related to the sampling criteria, which can happen if researchers rely on convenience techniques, such as snowball sampling or drop respondents that fail attention checks.

In the presence of endogenous sampling, unweighted estimates may be biased, but we can correct that bias when participants are up-weighted "by the inverse of the compliance score, then performing IV estimation" (Aronow and Carnegie 2013, 498). This process still leverages "the random assignment of the instrument to achieve a consistent estimator of the ATE for compliers", while the sample of compliers also has "a covariate distribution that matches that of the full population" (493). I typically recommend against this in the manuscript, however, because we must assume that inattentive participants will behave like attentive participants that are demographically similar to them (Alvarez et al. 2019).

The more fundamental reason why we use weights in the manuscript is to implicitly

[5]This is slightly different than Berinsky et al. (2019) who try to identify average partial effects in the presence of unmodeled effect heterogeneity, which interaction terms are more appropriate to handle (Solon, Haider and Wooldridge 2015).

state that we believe inattentive respondents are from a population whose variance is larger that the population variance for attentive respondents. In other words, less faith is put in the precision of the measurement for less attentive respondents and more faith in the precision of attentive ones. Under endogenous sampling, the ordinary and weighted linear regression results should diverge because they have different probability limits. If there is no endogenous sampling, the results should be similar between the two models. In conjunction with simulating the LATE, weighted regression allows researchers to highlight more precisely how the average treatment effect among the population they wish to generalize to differs with regard to attentive and inattentive individuals.

Weighting does have some drawbacks, however, one of which concerns statistical power. If researchers are concerned that they have too few observations to employ the techniques outlined in the manuscript and they have a treatment effect size in mind that is informed from the literature, they can perform a power calculation to see if they still have a sufficient number of effective observations to likely record a treatment effect if one exists. This functionality is offered in the R package. Another approach is to merely up-weight instead by using the inverse of respondents' average attention $\left(\frac{1}{\sum_{i=1}^{n} 1-s_i \frac{1}{n}}\right)$. There is not, however, a substantial difference between up- versus down-weighting.

For instance, let us compare two participants under the two weighting schemes, the first is very dissimilar (far) from and the second is very similar (close) to the text that they read. If the two respondents had average attentions ($s_i \frac{1}{n}$) of 0.9 (very far) and 0.2 (very close), they would score 0.27 ($1 - 0.9^3$) and 0.99 ($1 - 0.2^3$) under the initial $k$=3 weighting approach described in the manuscript (remember, we down-weight or penalize individuals for low attention), and their weights would be 1.11 (1/0.9) and 5 (1/0.2) using

the inverse of their average attention (now, we up-weight based on attention). This is relatively the same weighting magnitude of high attention to low attention participants ($0.99/0.27 = 3.67$ attentive to inattentive respondents versus $5/1.\bar{1} \approx 4.5:1$).

So, the first major difference is the magnitude of potential impact that low and high attentive respondents have on the treatment effect, with higher attention individuals receiving a higher magnitude of weight using the inverse of their average attention (though this magnitude could be adjusted by $k$). Second, and more important, I do not recommend up-weighting because the certainty around our point estimates of the treatment effect will automatically be smaller (for the same reasons why down-weighting reduces our effective number of observations). Our smaller bounds of uncertainty, therefore, are not because we have more information and I prefer to maintain a higher degree of uncertainty as a trade off for statistical power if possible.

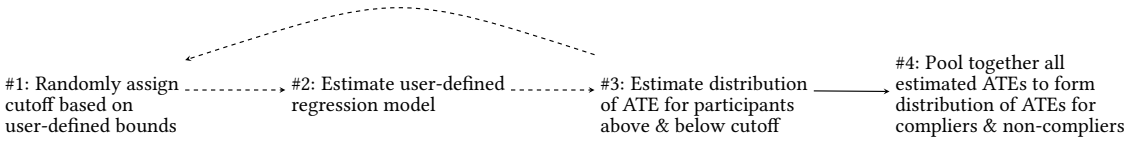
Another way that researchers can adjust how severely inattentive participants are down-weighted in comparison to attentive respondents is by varying $k$. The motivating determinants I use in the manuscript to set $k$ are (1) how much weight low attentive participants are given, and (2) how highly the average similarity measures are correlated with the "correct" answer as determined by a human. I show in Section A.5 how the results in the manuscript compare using different values of $k$.

## SM.4 Simulating the Treatment Effect for Compliers

I visually outline in Figure SM.3 the process of estimating the distribution of average treatment effects among participants that likely received the treatment. The first step of each round is to randomly assign the cutoff threshold, such that participants under this threshold are considered "non-compliers" and participants above are labeled as

"compliers". The cutoffs used in the manuscript, for instance, were drawn from a uniform distribution and varied randomly between 0 and 0.1, which does not correspond

**Figure SM.3:** Process of simulating the distribution of the average treatment effect among compliers and non-compliers.

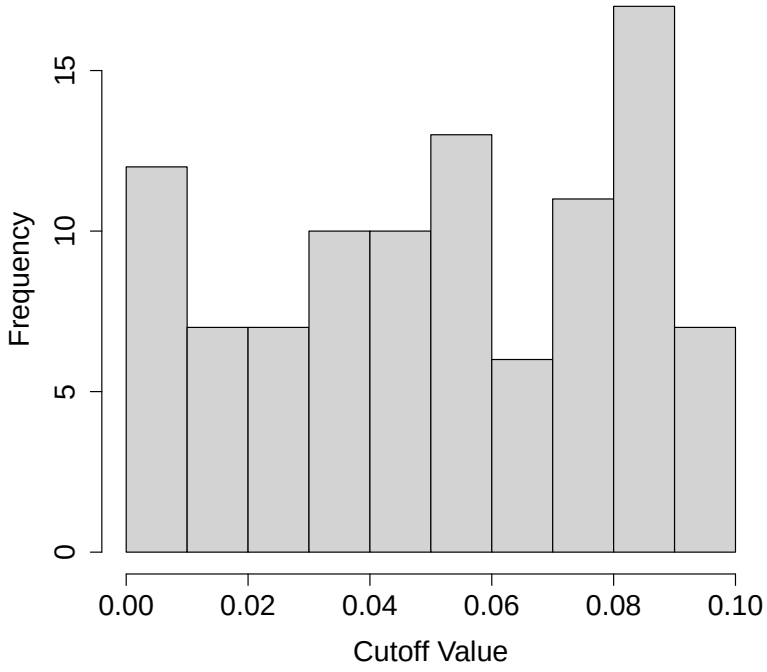



*Notes:* The dotted arrows connect stages of the simulation that are repeated each round. The solid arrow connects the stages that are repeated with the final output.

to the same percentage of respondents that would have failed the manipulation check in Kane (2020) because so few respondents passed based on human coders' assessments. The average number of respondents labeled as "non-compliers" was 170 throughout the simulations, for instance, while 423 respondents would be removed by list-wise deletion based on correctness. If nothing else, this should decrease the precision of the ATE of compliers because we are labeling more inattentive individuals as compliers (which it does not). The distribution of cutoffs that were used in the manuscript, which includes 100 simulations in total, is shown in Figure SM.4.

Once the cutoff is assigned at the beginning of each round, we run the user-defined regression model (Stage #2). In the manuscript, the outcome indicated whether a respondent selected the new story about President Trump (0=no, 1=yes) and the predictors were an interaction between party identification and the treatment. From this regression model, we can estimate the average treatment effect among those participants that we labeled as compliers and the ATE for non-compliers (Stage #3). We then store the distributions that are estimated for each group and repeat Stages 1 through 3 for a

**Figure SM.4:** Distribution of cutoffs to distinguish compliers and non-compliers for simulations of ATE distributions.




sufficient number of iterations.

I recommend completing at least 100 iterations to adequately sample around the cutoff space, especially if the cutoff is higher and there is more space to cover. I advise starting at 100 iterations to get a feel for how long it takes to compute and to get the proper sampling area for the cutoff. Then, researchers can increase the number of simulated rounds to 1,000 or even 10,000 for their final estimates. There are diminishing computational returns and there is very little difference substantively or statistically between using 100 or 10,000 iterations. For instance, there was no substantive difference in the results for the examples in Section A.6, but the additional 9,900 iterations took over 2 hours to complete on a typical laptop. After a sufficient number of simulation

rounds, we can investigate the pooled distribution of the all of the marginal treatment effects. The resulting distributions, for instance, are shown in Figure 4 in the manuscript.

The simulation process mirrors an instrumental variable approach in which we estimate the average effect of the treatment among the whole population of compliers. To illustrate, let us first consider when both the treatment and the instrument are binary, we can estimate the local average treatment effect as:

$$E(Y_{i1} - Y_{i0}|D_{i0} = 0, D_{i1} = 1) = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{P[D_i = 1|Z_i = 1] - P[D_i = 1|Z_i = 0]} \tag{1}$$

The conditions for identifying the LATE when our instrument is continuous are similar to the binary case, but we have to make the additional assumption of strict monotonicity. In other words, if our instrument $Z_i$ has a finite support and takes values from 0, ..., $J$, (in this case $J = 1$) the higher a participant's value of attention, the higher the probability that they received the treatment, $P(D_i = 1|Z_i = j) > P(D_i = 1|Z_i = j - 1)$. So, we can estimate the LATE if we do many pairwise comparisons between the compliers (group $j$) with non-compliers (group $j - 1$) varying who is a complier, which is why monotonicity is needed. This means that we estimate an ATE that is equal to the average effect of the treatment among the whole population of compliers and non-compliers. Put differently, we estimate the LATE using the average of ATEs from each complier subgroup.

Still, and most importantly, we can characterize the sampling distribution of both the complier and non-compliers, which we do not easily get if we use an instrumental variable approach with a two-stage regression model. Another key difference is that

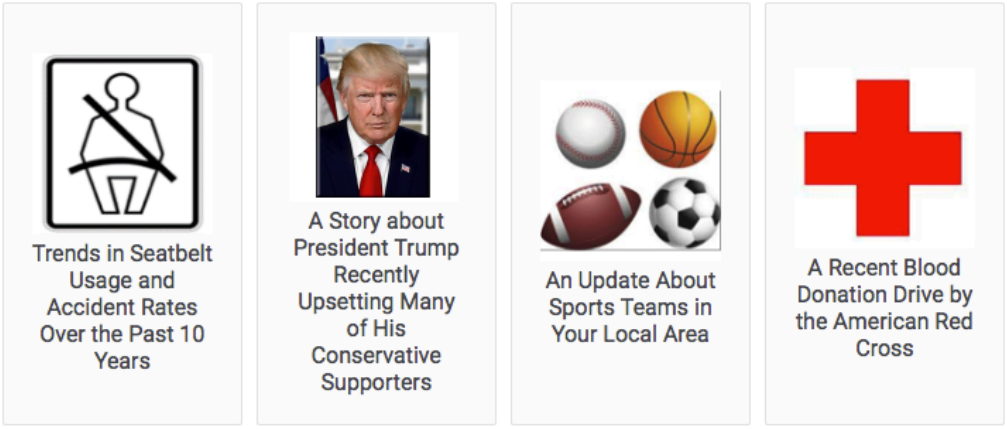
when we estimate a two-stage regression model it does not yield the exact same result because it uses a weighted average of Wald ratios, which counts some sub-groups more often than others. Our sampling approach versus the two-stage regression model should only produce the same LATE if the treatment effect is the same among all complier sub-groups. Still, I show the traditional instrumental variable approach yields comparable results to our simulations in Section SM.5. I prefer the simulation approach in the manuscript because we can investigate the sampling distribution of compliers and non-compliers, rather than only the treatment effect of compliers on average. Nonetheless, these two approaches are more desirable to a closed-ended or human coded open-ended manipulation checks that force one, specific, arbitrary threshold of correctness.

## SM.5 Re-analysis of Kane (2020)

In this section, I replicate the main tables and figures that are central to the findings of the second study in Kane (2020). I also provide all of the supporting evidence for the extensions that I mentioned in the manuscript, including how to investigate the predictors of attention, as well as how to select $k$.

The experiment in Kane (2020) manipulated the content of the news story about President Trump, seen in Figure SM.5, to explore how partisans select media based on the political content of the headline. After respondents viewed these news stories they were asked: "If you had to pick one, which of the following news stories would you want to read?". Subsequently, participants were asked to recall what the news story pertaining to Trump stated to confirm that participants actually read the headline and retained the information.

**Figure SM.5:** Experimental image condition from Kane (2019, A14).




To re-analyze the results, Table SM.1 begins by highlighting the frequency, mean attention, and mean outcome response of participants assigned to each treatment condition by party ID. There does not immediately appear to be any substantively differential assignment to treatment conditions by party ID. Nor, does it appear that partisans' average attention is associated with the treatment condition they are assigned to or their propensity to select the news story about President Trump. Nevertheless, I further investigate the correlation between attention, party ID, and treatment more formally below.

I also replicate Figure G1 in the Appendix of Kane 2019 (A23) in Table SM.2, which displays respondents' original correctness classification by the human coders. In general, this is concerning for practitioners that use human coders because it is often difficult to assess whether an open-ended response is an accurate representation of the prompt.

Next, Table SM.3 shows the estimated coefficients from a logistic regression in which the outcome indicates whether participants selected the news story about President Trump (1) or any of the other three news story options (0). The "United"

**Table SM.1:** Frequency, mean attention, and mean likelihood of selecting the news story about the President by treatment condition and party identification.

| Treatment | Party ID | $\bar{x}_{\text{Attention}}$ | $\bar{x}_{\text{Select Trump Story}}$ | N |
|---|---|---|---|---|
| Control | Independent | 0.36 | 0.29 | 62 |
| Control | Democrat | 0.35 | 0.19 | 96 |
| Control | Republican | 0.36 | 0.34 | 77 |
| Disunited | Independent | 0.34 | 0.25 | 84 |
| Disunited | Democrat | 0.32 | 0.38 | 86 |
| Disunited | Republican | 0.47 | 0.38 | 74 |
| United | Independent | 0.40 | 0.19 | 80 |
| United | Democrat | 0.39 | 0.26 | 95 |
| United | Republican | 0.33 | 0.47 | 88 |

**Table SM.2:** Replication of Figure G1, "Factual Manipulation Check Results".

| | Control | Disunited | United |
|---|---|---|---|
| Correct | 0.481 | 0.426 | 0.388 |
| Incorrect | 0.519 | 0.574 | 0.612 |

*Notes:* Proportion of respondents that answered the manipulation check "correctly" by treatment. Footnote from original table: "Qualtrics data. Diagonal indicates that factual manipulation check (FMC) responses vary systematically with treatment assignment ($\chi 2$ (631.99); p<.001). Cramér's V, a measure of association between categorical variables, is equal to 0.653, indicating a substantively strong association between the variables."

condition depicts Trump's conservative supporters as being pleased with him, while the "Disunited" condition depicts Trump's conservative supporters as being displeased with him. The control condition features basic information about President Trump. Independents and the control treatment are the two baseline categories to which effects should be compared.

**Table SM.3:** Estimated coefficients from base model (interaction by treatment and party ID).

| | |
|---|---|
| Disunited | −0.20 |
| | (0.38) |
| United | −0.57 |
| | (0.40) |
| Democrat | −0.57 |
| | (0.38) |
| Republican | 0.22 |
| | (0.37) |
| Disunited:Democrat | 1.20* |
| | (0.51) |
| United:Democrat | 1.01 |
| | (0.53) |
| Disunited:Republican | 0.38 |
| | (0.51) |
| United:Republican | 1.11* |
| | (0.51) |
| Constant | −0.89** |
| | (0.28) |

*Notes:* N=742, standard errors are presented in the parentheses. P-values are based on two-tailed hypothesis tests, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The results in Table SM.3 suggest that, in comparison to Independents that view the control textual treatment, Democrats prefer the disunited partisan story. Republicans, on the other hand, are only more likely to select the news story about a United Republican

party compared to Independents assigned to the control. However, these results do not estimate the marginal effect of moving from the Disunited treatment from the control treatment, for instance. This is the primary reason why I present the marginal effects in the manuscript. This approach also more closely mimics the analysis found in Table F1 in the Appendix of Kane, which reduces the sample to only Democrats or Republicans and regresses respondents' treatment on whether they selected the news story for each respective partisan group.

However, our goal is to see how the results change if we consider participants' attention. I report the estimated regression models that are used in the manuscript to create Figure 2 in in Table SM.4. The key takeaway from the weighted models in the manuscript is that participants assigned to the "Control to Disunited" and "Disunited to United", regardless of whether they are Republican or Democrat, likely have a non-zero treatment effect. This is difficult to glean from Table SM.4, which is why I calculate the marginal treatment effect and display it in Figure 3 of the manuscript.

We do not know, however, whether the treatment effects that we estimate across models are statistically differentiable from each other. In other words, is the estimated ATE of Democrats going from the "Control" to "Disunited" condition different based on whether we down-weight based on attention or keep the full sample? Researchers using the `openEnded` package can investigate the difference between weighting options using the `plotDifferences` function. In our application, there is little divergence between the ATEs estimated by the three weighting schemes. Interestingly, when we examine the model fit in the bottom of Table SM.4, the down-weighted model in the third column has the best model fit though it has the fewest number of observations.

**Table SM.4:** Full Estimated Coefficients for Figure 3 in the Manuscript.

| | Unweighted | List-Wise Deletion | Weighted |
|---|---|---|---|
| Treatment$_{Disunited}$ | −0.205 | 0.383 | 1.705* |
| | (0.377) | (0.537) | (0.870) |
| Treatment$_{United}$ | −0.573 | −0.105 | 0.916 |
| | (0.400) | (0.543) | (0.851) |
| Democrat | −0.573 | −1.235 | −0.223 |
| | (0.383) | (0.650) | (0.931) |
| Republican | 0.220 | 0.288 | 1.386 |
| | (0.369) | (0.495) | (0.838) |
| Democrat:Treatment$_{Disunited}$ | 1.197* | 1.788* | 0.453 |
| | (0.509) | (0.831) | (1.088) |
| Democrat:Treatment$_{United}$ | 1.009* | 0.963 | 0.164 |
| | (0.532) | (0.858) | (1.065) |
| Republican:Treatment$_{Disunited}$ | 0.382 | −0.488 | −2.255* |
| | (0.507) | (0.706) | (1.008) |
| Republican:Treatment$_{United}$ | 1.110* | 0.999 | 0.131 |
| | (0.514) | (0.723) | (0.998) |
| Constant | −0.894** | −0.875* | −1.792* |
| | (0.280) | (0.376) | (0.764) |
| AIC | 899.299 | 398.769 | 339.237 |
| BIC | 940.784 | 432.656 | 371.689 |
| Log Likelihood | −440.650 | −190.385 | −160.619 |
| N | 742 | 319 | 272 |

*Notes:* Total N=742, standard errors are presented in the parentheses. Statiscal reliability is reported as $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

This is further evidence that inattentive participants are contributing additional noise to our model.

A vital consideration for researchers when deciding upon the "correct" model is what value to set $k$. I selected $k = 3$ in the manuscript because I wanted to more heavily discount inattentive participants, and because there are diminishing returns for increased values of $k$. Figure SM.6 shows that when $k = [3,5]$ the correlation between respondents' measure of similarity and the "correct" answer was over 0.76. Researchers can create a similar figure using the `plotK` function in package. Unsurprisingly, as $k$ increases, the overall treatment effects get pulled toward zero because there are fewer and fewer observations.

We also want to inspect which participants are more likely to be attentive and whether partisans are more likely to be (in)attentive to treatment conditions they are prone to (dis)like. First, Table SM.5 highlights that participants who are older, non-Hispanic White, women, or have a college degree are more likely, on average, to provide a response that is similar to the text that they read. If researchers are worried that these biases will be reflected in their estimation of the PATE and LATE, I advise readers to follow Aronow and Carnegie (2013) and up-weight inattentive participants so that the sample of compliers also has "a covariate distribution that matches that of the full population" (493).

Second, we can empirically verify in Table SM.5 that our assumption of non-differential attention by treatment and partisanship is held. The bottom half of Table SM.5 shows that respondents do not provide more or less attention based on their partisanship and treatment condition. This is important because we can at least demonstrate that

participants are not systematically assigned to a treatment they are prone to (dis)like (Table SM.1), nor that participants pay more or less attention based on which text they view as part of the treatment (Table SM.5).

Lastly, to show that simulating the sampling distribution of the LATE retrieves a similar point estimate to a more traditional two-staged approach, Table SM.6 emphasizes the same main conclusion as Figure 4 in the manuscript: once we account for the likelihood that participants received the treatment, partisans were more likely to select stories they are prone to favor.

**Figure SM.6:** Correlation between participants' average similarity measure and the correct answer as determined by a human coder.

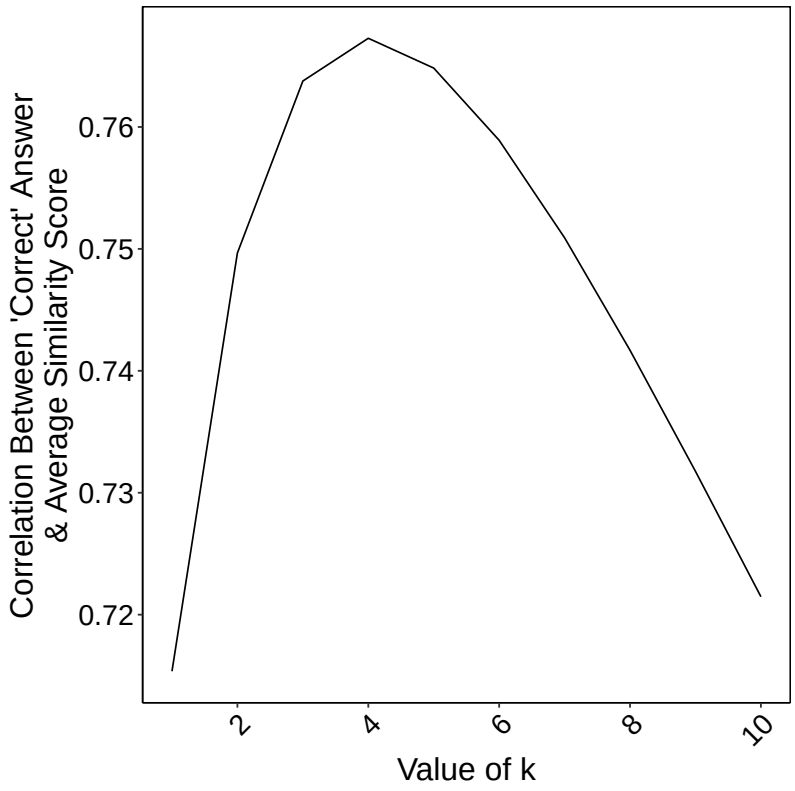


**Table SM.5:** Predicting attention using socio-demographic variables, treatment group, and partisanship.

| | Attention (Average Similarity Measure) | Attention (Human Correctness) |
|---|---|---|
| ***Socio-Demographic Factors*** | | |
| $\text{Age}_{(42,66]}$ | $0.062^{*}$ | $0.091^{*}$ |
| | (0.027) | (0.040) |
| $\text{Age}_{(66,90]}$ | $0.147^{***}$ | $0.218^{***}$ |
| | (0.038) | (0.057) |
| College Grad | $0.068^{*}$ | $0.083^{*}$ |
| | (0.027) | (0.041) |
| Non-White | $-0.092^{***}$ | $-0.121^{**}$ |
| | (0.026) | (0.038) |
| Income | −0.012 | −0.004 |
| | (0.009) | (0.013) |
| Male | $-0.058^{*}$ | $-0.092^{*}$ |
| | (0.027) | (0.041) |
| ***Political Factors*** | | |
| Democrat | 0.017 | −0.125 |
| | (0.053) | (0.080) |
| Republican | −0.005 | −0.020 |
| | (0.056) | (0.083) |
| $\text{Treatment}_{Disunited}$ | 0.012 | −0.157 |
| | (0.054) | (0.081) |
| $\text{Treatment}_{United}$ | 0.047 | −0.123 |
| | (0.054) | (0.082) |
| $\text{Democrat:Treatment}_{Disunited}$ | −0.037 | 0.150 |
| | (0.072) | (0.109) |
| $\text{Republican:Treatment}_{Disunited}$ | 0.076 | 0.150 |
| | (0.076) | (0.114) |
| $\text{Democrat:Treatment}_{United}$ | −0.009 | 0.109 |
| | (0.072) | (0.108) |
| $\text{Republican:Treatment}_{United}$ | −0.065 | −0.035 |
| | (0.074) | (0.111) |
| Constant | $0.374^{***}$ | $0.536^{***}$ |
| | (0.046) | (0.069) |
| AIC | 435.517 | 1037.351 |
| BIC | 509.267 | 1111.101 |
| Log Likelihood | −201.759 | −502.676 |

*Notes:* N=742, standard errors are presented in the parentheses. Statiscal reliability is reported as $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

**Table SM.6:** Second of two-staged regression model using attention as indicator of probability of receiving the treatment.

| | Select Trump Story |
|---|---|
| $\text{Treatment}_{Disunited}$ | −0.040 |
| | (0.076) |
| $\text{Treatment}_{United}$ | −0.103 |
| | (0.077) |
| Democrat | −0.103 |
| | (0.074) |
| Republican | 0.047 |
| | (0.077) |
| $\text{Democrat:Treatment}_{Disunited}$ | $0.237^{*}$ |
| | (0.101) |
| $\text{Democrat:Treatment}_{United}$ | 0.178 |
| | (0.101) |
| $\text{Republican:Treatment}_{Disunited}$ | 0.081 |
| | (0.106) |
| $\text{Republican:Treatment}_{United}$ | $0.231^{*}$ |
| | (0.104) |
| Constant | $0.290^{***}$ |
| | (0.058) |
| $R^2$ | 0.039 |

*Notes:* N=742, standard errors are presented in the parentheses. Statiscal reliability is reported as $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

## SM.6 Implementation in R and Additional Application

The second example I present of an open-ended manipulation check comes from a self-designed and implemented study, which investigated rhetorical responsiveness in the Catholic Church and the motivations for the Pope to be responsive. In this section, I describe the survey design and show how to analyze the results using the R package that I created for open-ended manipulation checks. Please consult the package's GitHub page for the most up-to-date references and vignettes as the functionality of the package improves.

The survey experiments, which were conducted in Brazil and Mexico (N≈5,000), assessed how members react when the Pope, a formally unaccountable leader, provides responsiveness in his rhetoric. The primary implication of the theory and findings is that members react positively and provide the Church with their support when the Pope discusses issues that are salient to Catholics. Members' existing support, furthermore, conditions the impact of responsiveness or non-responsiveness, such that regular church attendees drive this relationship in the aggregate.

Participants of the online survey experiments were limited to self-identified Catholics. The survey was carried out among a nationally representative quota sample from each Brazil and Mexico (N≈2,500) and administered online by the international polling firm Respondi. Respondi employs a combination of online and offline recruitment methods to ensure that the panels can be used for conducting representative surveys. The two samples were nationally representative by age, gender, and region derived from population censuses to ensure that the sample margins match those in the target population.

Respondents were presented with three selected news headlines on the same topic outlining recent statements made by the Pope (conflict, human rights, socio-political issues, economy, and control/religious issues). The three news headlines associated with each of the five topics are found in Table SM.7. These messages represent the typical language content and phrasing used in the media when describing the Pope's statements.

Respondents were randomly assigned to receive news stories pertaining to either (1) a topic that they believed is most important (the "responsive" treatment), or (2) one of the four other issue areas ("non-responsive"). Within those respondents that received "non-responsive" messages, there was an even probability of assignment to each topic. The ordering of questions, including treatment assignment, are shown in Figure SM.7.

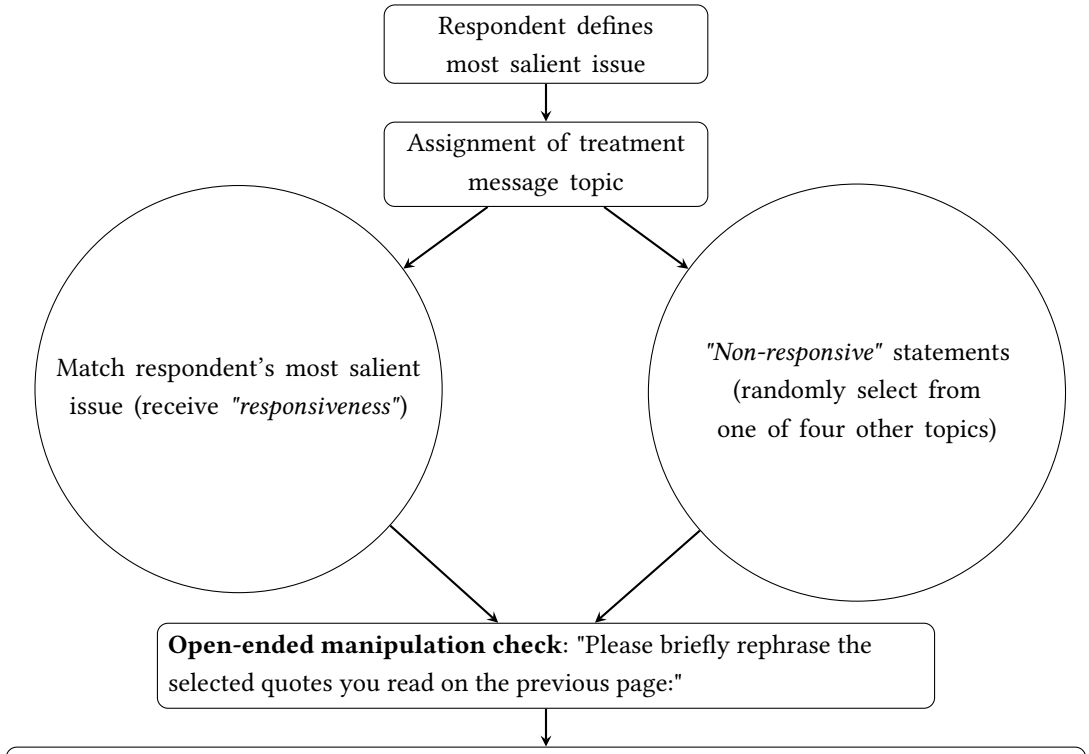
Before respondents viewed the textual treatment they were asked pre-treatment questions about their age, gender, region of residence, and political preferences related to the issues that were mentioned in the news treatments. Prior to the outcome questions, but after the textual treatment, participants were asked to recall the stories they read on the previous page in an open-ended response manipulation check. Afterward, respondents then expressed the degree to which they thought the Church is responsive, the degree to which they trusted the Church, and the degree to which they anticipated increasing their organizational participation.

**Table SM.7:** News headlines summarizing papal rhetoric for each issue area.

| |
|---|
| **Conflict** |
| 1. "Pope pleads for end to 'homicidal madness' of terrorism". |
| 2. "Pope meets with Colombian leaders in wake of peace deal". |
| 3. "Let's unite against war and violence, Pope urges at Roman synagogue". |
| **Economy** |
| 1. "Pope says economy must fight 'throwaway culture'". |
| 2. "Generate new models of economic progress, Pope urges business leaders". |
| 3. "'Economy of exclusion, inequality caused growth of poverty', says Pope". |
| **Socio-political issues** |
| 1. "Education and play are key to childhood, Pope tells Cuba, US youth". |
| 2. "Holy See backs global health goals, says 'leave no one behind'". |
| 3. "Pope asks: give immigrants compassion, not blame". |
| **Human rights** |
| 1. "Vatican diplomacy zeros-in on human rights in Africa". |
| 2. "For Pope, it's imperative: religious liberty is a gift from God. Defend it". |
| 3. "Pope says promotion of human rights is central to the commitment of the European Union". |
| **Control (neutral)** |
| 1. "Pope marks 80th birthday in Rome, addresses Cardinals at Mass". |
| 2. "If you're tempted to gossip, 'bite your tongue,' Pope says". |
| 3. "Love God now - because you might not have tomorrow, Pope says". |

*Notes*: The survey was translated from English to Spanish (for Mexican respondents) and Brazilian Portuguese (for Brazilian respondents).

**Figure SM.7:** Respondent assignment to treatment and outcome responses for survey experiment of Catholics.




- Outcome questions: "Please indicate how strongly you agree or disagree with the following statement from 1 (Strongly disagree) to 10 (Strongly agree):"
  - **Outcome 1:** "The Church is responsive to its members' needs and concerns."
  - **Outcome 2:** "I trust the Church."
  - **Outcome 3:** "I plan to attend more church services in the future."
  - **Outcome 4:** "I want to volunteer through the Church more in the future."
- **Outcome 5:** "Would you like to learn more about a political petition related to '[most preferred issue]'?"

  Please provide your answer below where 1 means you are not at all interested and 10 means that you are very interested."

As a visual reference, the distributions for the $n$-gram similarity measures (Jaccard and cosine) in each country are shown in Figure SM.8 and SM.9. To create these figures, we first need to download and install the library. All of the documentation for the functions and arguments included in the R package can be found on the GitHub webpage. Please consult the replication materials for full installation details. You can download the package by executing:

```
devtools::install_github('jeffreyziegler/openEnded', force=T)
```

Next, users can load in their data and specify which vector within their dataframe contains the prompts and which vector contains the responses to calculate their similarity measures. I import the data for the survey experiments in Brazil and Mexico, which are contained within the package. The vector of responses to the open-ended manipulation check are stored in `zieglerData$validityCheck` and the treatments that respondents read are stored in `zieglerData$textViewed`. To create our various $n$-gram similarity measures, such as the Jaccard and the cosine of the angle between the vectors, we can execute the function `similarityMeasures` as seen below. We assign $n$=3 as we did in the manuscript.

```
zieglerData <- similarityMeasures(dataframe=zieglerData,
                                  n_gram_measures_to_calculate=c("jaccard",
                                                                 "cosine",
                                                                 "jw", "dl"),
                                  prompt="textViewed",
                                  response="validityCheck",
                                  ngrams=3)
```

With our similarity measures in hand, we can plot the distribution of all respondents with the function `plotMeasures`. Figure SM.8, specifically, shows the plotted output from the code below. The default `plotSimilarity` does not currently include the ability to label select responses as seen in the manuscript.

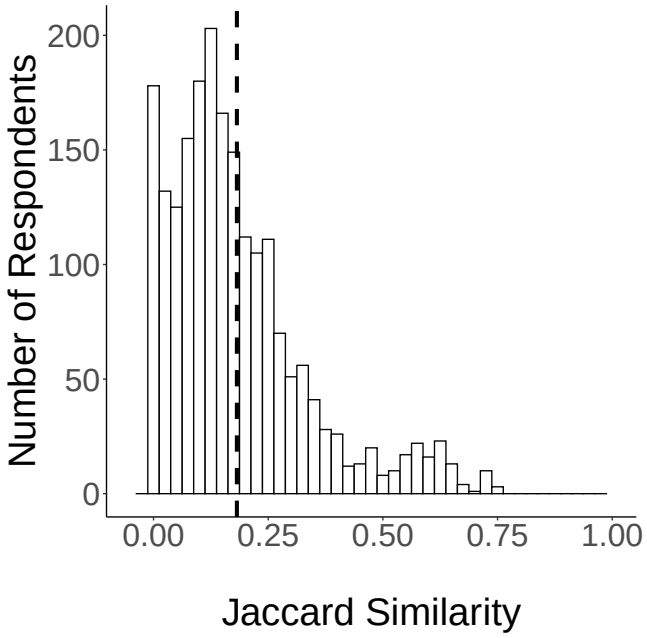
```
plotSimilarity(dataframe=zieglerData[which(zieglerData$Country=="Brazil"),],
               measure="jaccardSimilarity",
               plot_path="../figures/FigSM8a.pdf"
)
```

The distributions, especially in Mexico, are more highly skewed to the left than the data presented in the manuscript from Kane (2020), which means that more respondents will be down-weighted with low values of $k$. Nevertheless, the Jaccard and cosine measures are high correlated, as seen in Figure SM.10, which can be created with the function `plotSimilarityCorr`.

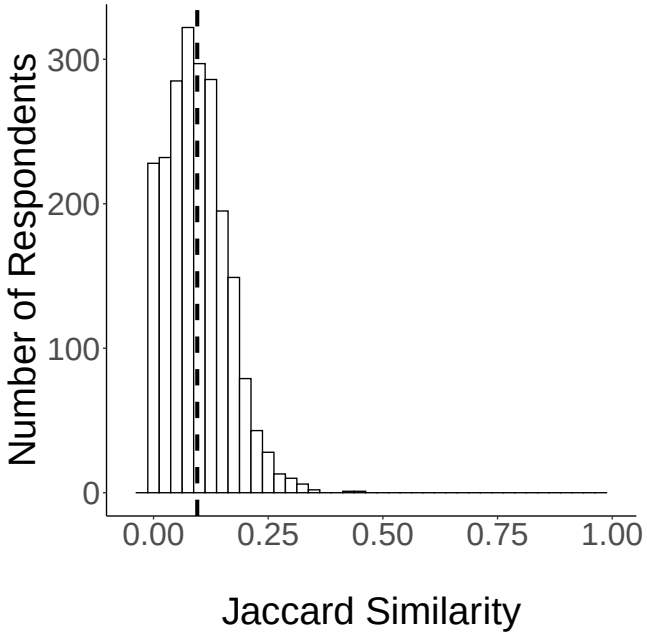
```
plotSimilarityCorr(dataframe=zieglerData,
                   measures=c("jaccardSimilarity",
                              "cosineSimilarity",
                              "jwSimilarity",
                              "dlSimilarity"),
                   labels=c("Jaccard (n-gram)",
                            "Cosine (n-gram)",
                            "Jaro (n-gram)",
                            "Damerau-Levenshtein (n-gram)"))
```

**Figure SM.8:** Distribution of raw Jaccard similarity measures for respondents in Brazil and Mexico.

**(a)** Brazil



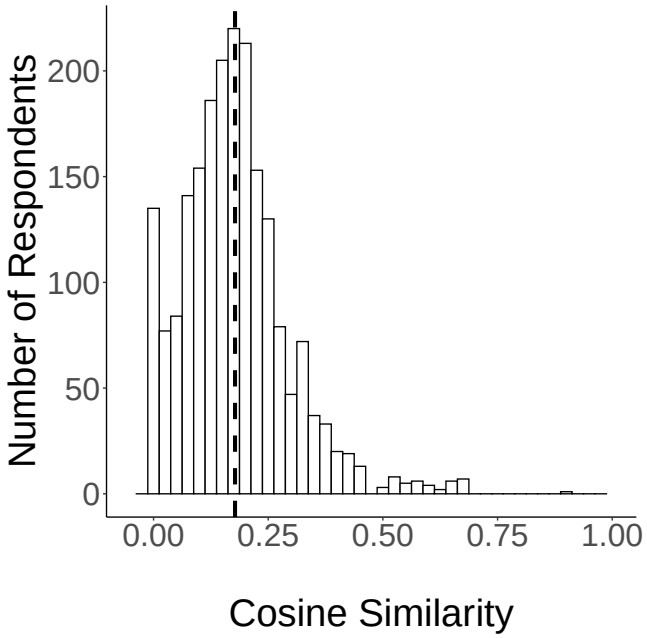

**(b)** Mexico



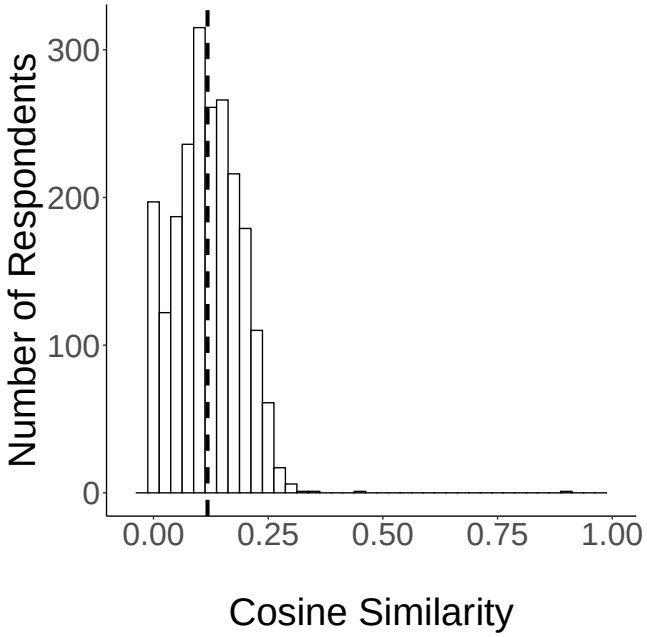

*Notes*: The mean distance for each country is represented by the vertical dotted-line.

**Figure SM.9:** Distribution of raw cosine of angles for respondents in Brazil and Mexico.

**(a)** Brazil

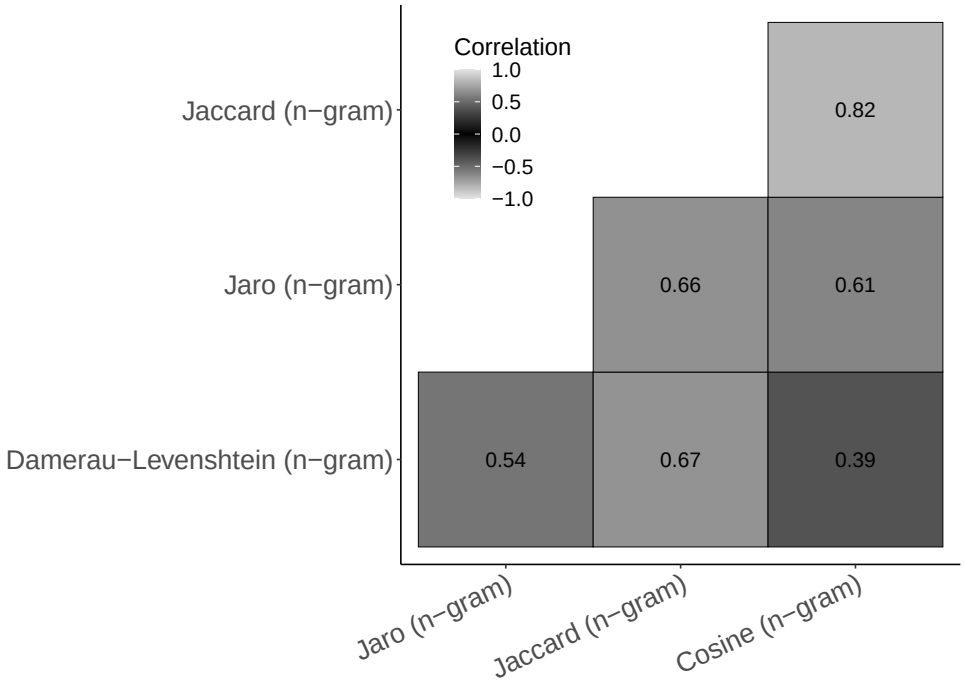



**(b)** Mexico




*Notes*: The mean distance for each country is represented by the vertical dotted-line.

**Figure SM.10:** Correlation between distance measures for respondents in Brazil and Mexico.




Now, I present the regression results from models estimated with (1) the full sample irrespective of attention, (2) a reduced sample using list-wise deletion based on an arbitrary threshold set for participants that "passed" (those respondents with weights ≥ 0.1) since I did not have human coders assess correctness, and (3) a weighted least squares model based on the weighted average of the Jaccard and cosine similarity measures.

To execute the three regressions, we can run the function `regressionComparison`, which estimates the three separate regression models. You do not need to calculate the average similarity, the function computes this for you, you only need to define a value for $k$ and which similarity measures to include in the averaged measure. The output of the regression models from this function will be automatically loaded into your

**Table SM.8:** Estimated coefficients from (1) regression with all observations, (2) weighted regression based on attentiveness, (3) regression on subsetted sample based on attentiveness.

| | *Outcome:* | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) Trust | (2) Trust | (3) Trust | (4) Responsive | (5) Responsive | (6) Responsive | (7) Volunteer | (8) Volunteer | (9) Volunteer | (10) Attendance | (11) Attendance | (12) Attendance | (13) Petition | (14) Petition | (15) Petition |
| Responsive papal messaging | $-0.26^{*}$ | $-0.34^{*}$ | $-0.30^{*}$ | $-0.06$ | $-0.07$ | $-0.04$ | $-0.53^{***}$ | $-0.70^{***}$ | $-0.65^{***}$ | $-0.35^{**}$ | $-0.45^{***}$ | $-0.41^{**}$ | $-0.16$ | $-0.17$ | $-0.15$ |
| | (0.13) | (0.13) | (0.14) | (0.13) | (0.14) | (0.14) | (0.13) | (0.14) | (0.14) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) |
| Attendance (Monthly) | $0.87^{***}$ | $0.91^{***}$ | $0.91^{***}$ | $0.79^{***}$ | $0.88^{***}$ | $0.85^{***}$ | $1.32^{***}$ | $1.28^{***}$ | $1.31^{***}$ | $1.49^{***}$ | $1.51^{***}$ | $1.53^{***}$ | $0.58^{***}$ | $0.59^{***}$ | $0.66^{***}$ |
| | (0.12) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) | (0.14) | (0.12) | (0.13) | (0.13) | (0.12) | (0.13) | (0.13) |
| Attendance (Weekly) | $1.84^{***}$ | $1.86^{***}$ | $1.87^{***}$ | $1.62^{***}$ | $1.71^{***}$ | $1.69^{***}$ | $2.39^{***}$ | $2.38^{***}$ | $2.40^{***}$ | $2.29^{***}$ | $2.35^{***}$ | $2.32^{***}$ | $0.92^{***}$ | $0.94^{***}$ | $1.00^{***}$ |
| | (0.12) | (0.12) | (0.12) | (0.12) | (0.13) | (0.13) | (0.12) | (0.13) | (0.13) | (0.11) | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) |
| Responsiveness*Attendance (Monthly) | $0.43^{*}$ | $0.46^{*}$ | $0.46^{*}$ | 0.25 | 0.16 | 0.22 | $0.63^{***}$ | $0.76^{***}$ | $0.74^{***}$ | $0.47^{**}$ | $0.54^{**}$ | $0.51^{**}$ | 0.10 | 0.12 | 0.13 |
| | (0.18) | (0.18) | (0.19) | (0.18) | (0.19) | (0.19) | (0.18) | (0.19) | (0.19) | (0.17) | (0.18) | (0.18) | (0.18) | (0.18) | (0.19) |
| Responsiveness*Attendance (Weekly) | $0.45^{**}$ | $0.52^{**}$ | $0.49^{**}$ | $0.44^{**}$ | $0.36^{*}$ | $0.41^{*}$ | $0.72^{***}$ | $0.85^{***}$ | $0.81^{***}$ | $0.63^{***}$ | $0.69^{***}$ | $0.65^{***}$ | 0.30 | 0.26 | 0.26 |
| | (0.17) | (0.17) | (0.17) | (0.17) | (0.17) | (0.18) | (0.17) | (0.18) | (0.18) | (0.16) | (0.17) | (0.17) | (0.17) | (0.17) | (0.17) |
| Constant | $6.11^{***}$ | $6.15^{***}$ | $6.07^{***}$ | $5.45^{***}$ | $5.48^{***}$ | $5.38^{***}$ | $5.13^{***}$ | $5.29^{***}$ | $5.15^{***}$ | $5.44^{***}$ | $5.51^{***}$ | $5.43^{***}$ | $7.07^{***}$ | $7.19^{***}$ | $7.02^{***}$ |
| | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.10) | (0.09) | (0.10) | (0.10) | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) | (0.09) |
| Weights | | ✓ | | | ✓ | | | ✓ | | | ✓ | | | ✓ | |
| $R^2$ | 0.13 | 0.14 | 0.13 | 0.10 | 0.11 | 0.11 | 0.19 | 0.20 | 0.20 | 0.20 | 0.21 | 0.20 | 0.04 | 0.04 | 0.04 |
| Adj. $R^2$ | 0.13 | 0.14 | 0.13 | 0.10 | 0.11 | 0.11 | 0.19 | 0.20 | 0.20 | 0.20 | 0.21 | 0.20 | 0.04 | 0.04 | 0.04 |
| N | 4237 | 3971 | 3852 | 4237 | 3971 | 3852 | 4237 | 3971 | 3852 | 4237 | 3971 | 3852 | 4237 | 3971 | 3852 |

*Notes:* Standard errors are presented in the parentheses.

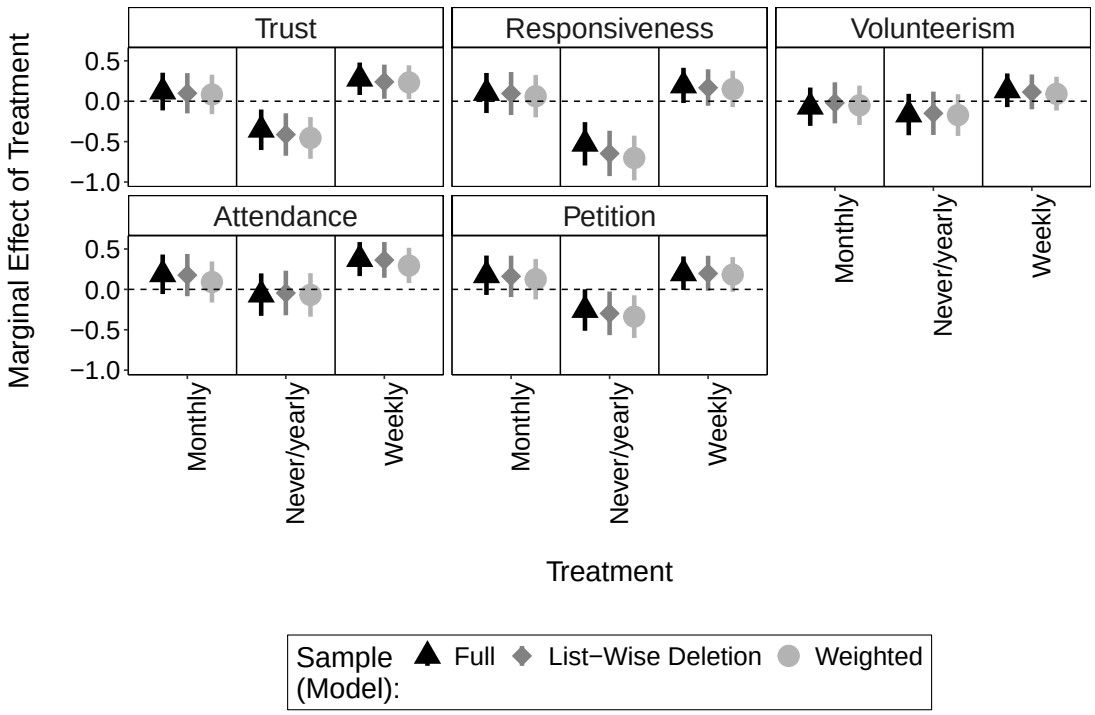
global environment (for instance, labeled as name of "`baseModel_`" or "`weightedModel_`" + outcome) and can be used as typical regression objects in R, so we can get the estimated coefficients to reproduce Table SM.8.

Once R has estimated our three regression models, the function also estimates and plots the average marginal effects with the function. An example of the output is seen in Figure SM.11, which indicates that dedicated members (those that attend church weekly) were more likely to increase their anticipated future attendance of Church services. When asked how strongly respondents agree with the statement, "I plan to attend more church services in the future", members that attended church services weekly were more likely to increase their support if they received responsiveness.

The estimated average treatment effect of receiving papal responsiveness for weekly attendees was associated with about a 0.3 point increase in the strength of their anticipated attendance of church services. These findings suggest that respondents' were more willing to view the Church as responsive, and more willing to participate in the Church, when they receive responsive papal statements. The results do not change substantively or statistically when the full sample is used versus samples that exclude or weight respondents based on attention. This may signal that inattentive participants and attentive participants do not respond to the outcomes systematically different, or at least not enough to alter the overall treatment effects.

To double-check whether attentive and inattentive participants respond differently in a systematic manner, which may explain some of the null estimates of the overall ATEs in Figure SM.11, I simulate the distribution of ATE for compliers and non-compliers. We can achieve this by executing `complierATE`, which will yield a plot similar to Figure SM.12.

**Figure SM.11:** Marginal treatment effects by church attendance and sample.

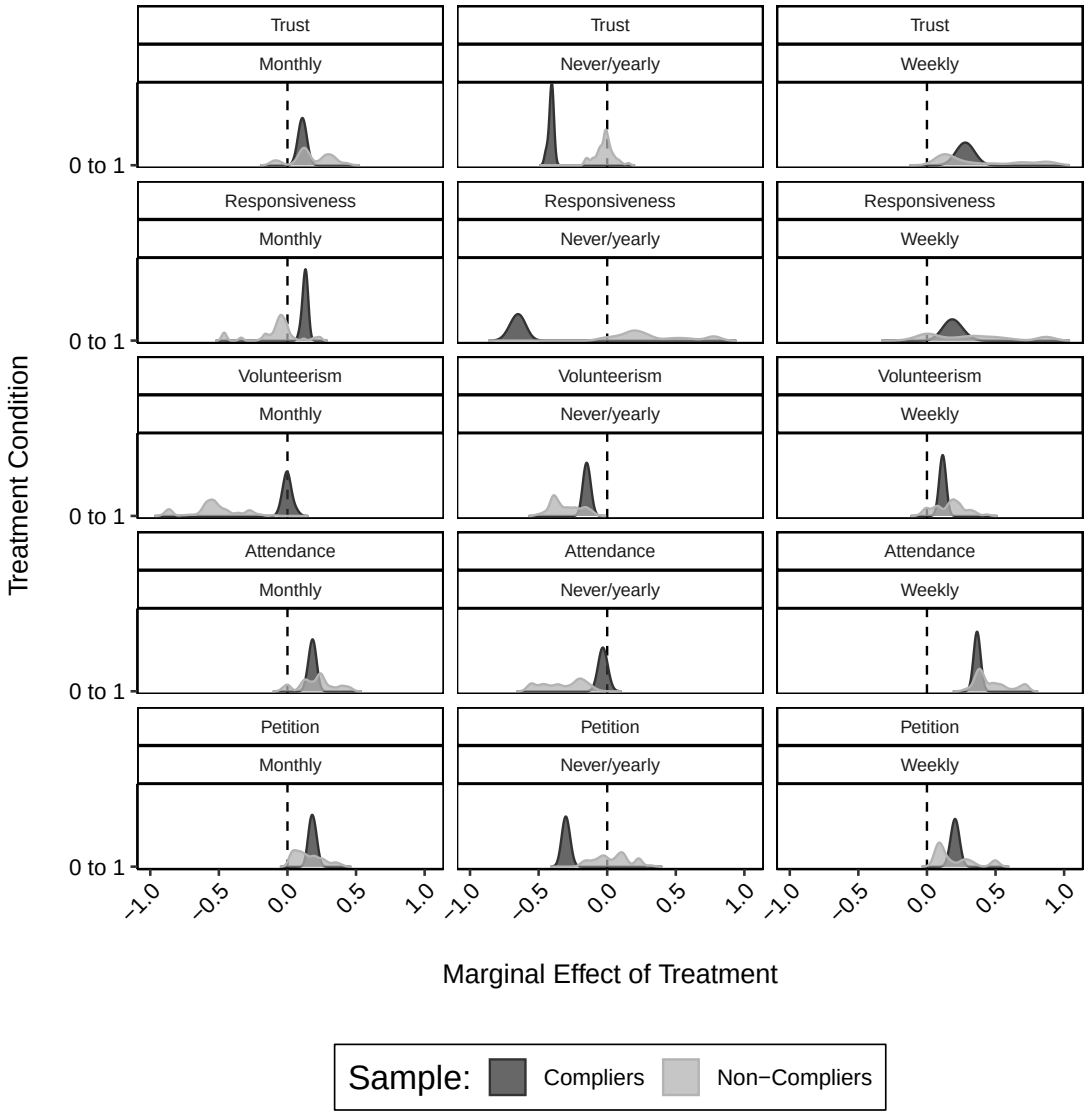



*Notes*: The figure plots marginal effect of the treatment measured by the change in the predicted level of support among the outcome categories. The mean marginal effects is represented by the solid point, while the 2.5%-97.5% percentiles of the sampling distributions are designated by the vertical lines. The marginal effects of each country are generated from 10,000 simulations that use asymptotic normal approximation to the log-likelihood to estimate the first difference for each category of attendance.

The user merely clarifies what the cutoff threshold which represents that maximum value of attention at which a participant would be considered a non-complier, and $n$ which references how many simulations the user wishes to perform (the default is 100 which matches the application in the manuscript).

Figure SM.12 plots the distribution of treatment effects for 100 simulations of the ATE for participants above (those that "passed") and below (those that "failed") a randomly selected weight threshold between 0 and 0.2. Beginning with those participants that

**Figure SM.12:** Distribution of average marginal treatment effects by church attendance for respondents that likely absorbed the treatment and those that did not.




*Notes*: The figure plots the median marginal effects of respondents that "passed" the manipulation check. The vertical lines represent the 2.5%-97.5% percentiles of the sampling distribution of the average marginal effect for compliers and non-compliers. Each distribution consists of $N = 100$.

would pass the manipulation check, we can see that the ATE typically increases as respondents' church attendance increases. Moreover, the distribution is tightly compact showing little variation in the ATE of compliers. Non-compliers do not consistently differ from compliers, with the exception of a few outcomes. Rather, non-compliers appear to add more uncertainty and heterogeneity into the average treatment effect, which may explain the lack of precision for the ATEs in Figure SM.11.

# References


Alvarez, R. Michael, Lonna Rae Atkeson, Ines Levin and Yimeng Li. 2019. "Paying attention to inattentive survey respondents." *Political Analysis* 27(2):145–162.

Aronow, Peter M. and Allison Carnegie. 2013. "Beyond LATE: Estimation of the average treatment effect with an instrumental variable." *Political Analysis* 21(4):492–506.

Banks, Antoine J. and Nicholas A. Valentino. 2012. "Emotional substrates of white racial attitudes." *American Journal of Political Science* 56(2):286–297.

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances and Christopher Warshaw. 2019. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?" *Political Science Research and Methods* pp. 1–8.

Bishop, George F. 1987. "Experiments with the middle response alternative in survey questions." *Public Opinion Quarterly* 51(2):220–232.

Brierley, Sarah, Eric Kramon and George Kwaku Ofosu. 2020. "The moderating effect of debates on political attitudes." *American Journal of Political Science* 64(1):19–37.

Clifford, Scott and Jennifer Jerit. 2014. "Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies." *Journal of Experimental Political Science* 1(2):120–131.

Denny, Matthew J. and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2):168–189.

Dietrich, Bryce J., Jeffery J. Mondak and Tarah Williams. 2020. "Using the Audio from Telephone Surveys for Political Science Research." *Working Paper* .

Edwards, Pearce and Daniel Arnon. 2019. "Violence on Many Sides: Framing Effects on Protest and Support for Repression." *British Journal of Political Science* pp. 1–19.

Foote, Jonathan T. 1997. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*. Vol. 3229 International Society for Optics and Photonics pp. 138–147.

Friedman, Ronald S. and Bárbara Sutton. 2013. "Selling the war? System-justifying effects of commercial advertising on civilian casualty tolerance." *Political Psychology* 34(3):351–367.

Garrett, Kristin N. and Joshua M. Jansa. 2015. "Interest group influence in policy diffusion networks." *State Politics & Policy Quarterly* 15(3):387–417.

Geer, John G. 1988. "What do open-ended questions measure?" *Public Opinion Quarterly* 52(3):365–367.

Geer, John G. 1991. "Do open-ended questions measure "salient" issues?" *Public Opinion Quarterly* 55(3):360–370.

Holland, Jennifer L and Leah Melani Christian. 2009. "The influence of topic interest and interactive probing on responses to open-ended questions in web surveys." *Social Science Computer Review* 27(2):196–212.

Hopkins, Daniel J. 2015. "The upside of accents: Language, inter-group difference, and attitudes toward immigration." *British Journal of Political Science* 45(3):531–557.

Iyengar, Shanto, Kyu S. Hahn, Jon A. Krosnick and John Walker. 2008. "Selective exposure to campaign communication: The role of anticipated agreement and issue public membership." *The Journal of Politics* 70(1):186–200.

Jamieson, Thomas and Nicholas Weller. 2019. "The Effects of Certain and Uncertain Incentives on Effort and Knowledge Accuracy." *Journal of Experimental Political Science* pp. 1–14.

Jansa, Joshua M., Eric R. Hansen and Virginia H. Gray. 2019. "Copy and paste lawmaking: legislative professionalism and policy reinvention in the states." *American Politics Research* 47(4):739–767.

Kane, John V. 2020. "Fight Clubs: Media Coverage of Party (Dis) unity and Citizens' Selective Exposure to It." *Political Research Quarterly* 73(2):276–292.

Kane, John V. and Jason Barabas. 2019. "No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments." *American Journal of Political Science* 63(1):234–249.

Keiser, Lael R. and Susan M. Miller. 2020. "Does Administrative Burden Influence Public Support for Government Programs? Evidence from a Survey Experiment." *Public Administration Review* 80(1):137–150.

Kim, Jeong Hyun and Yesola Kweon. 2020. "Status Threat and Opposition to Gender Equality Policies: Evidence from a Survey Experiment in South Korea." *Working Paper* .

Knox, Dean and Christopher Lucas. 2020. "A dynamic model of speech for the social sciences." *American Political Science Review* Conditionally Accepted.

Krosnick, Jon A. 1999. "Survey research." *Annual review of psychology* 50(1):537–567.

Kusner, Matt, Yu Sun, Nicholas Kolkin and Kilian Weinberger. 2015. From Word Embeddings to Document Distances. In *International conference on machine learning*. pp. 957–966.

Ladam, Christina. 2019. "Does Process Matter? Direct Democracy and Citizens' Perceptions of Laws." *Journal of Experimental Political Science* pp. 1–6.

McClendon, Gwyneth and Rachel Beatty Riedl. 2015. "Religion as a stimulant of political participation: Experimental evidence from Nairobi, Kenya." *The Journal of Politics* 77(4):1045–1057.

McCune, Bruce, James B. Grace and Dean L. Urban. 2002. *Analysis of ecological communities*. Vol. 28 MjM software design Gleneden Beach, OR.

Presser, Stanley and Howard Schuman. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Solon, Gary, Steven J. Haider and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human resources* 50(2):301–316.

Tourangeau, Roger, Lance J Rips and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press.

Van der Loo, Mark P.J. 2014. "The stringdist Package for Approximate String Matching." *The R Journal* 6(1):111–122.

Weber, Christopher and Matthew Thornton. 2012. "Courting Christians: How political candidates prime religious considerations in campaign ads." *The Journal of Politics* 74(2):400–413.

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the flow of policy ideas in legislatures: A text reuse approach." *American Journal of Political Science* 59(4):943–956.