A Text-As-Data Approach for Using Open-Ended Responses as Manipulation Checks

Jeffrey Ziegler[†]

Abstract

Participants that complete online surveys and experiments may be inattentive, which can hinder researchers' ability to draw substantive or causal inferences. As such, many practitioners include multiple factual or informational closed-ended manipulation checks to identify low-attention respondents. However, closed-ended manipulation checks are either correct or incorrect, which allows participants to more easily guess and it reduces the potential variation in attention between respondents. In response to these shortcomings, I develop an automatic and standardized methodology to measure attention that relies on the text that respondents provide in an open-ended manipulation check. There are multiple benefits to this approach. First, it provides a continuous measure of attention, which allows for greater variation between respondents. Second, it reduces the reliance on subjective, paid humans to analyze open-ended responses. Last, I outline how to diagnose the impact of inattentive workers on the overall results, including how to assess the average treatment effect of those respondents that likely received the treatment. I provide easy-to-use software in R to implement these suggestions for open-ended manipulation checks.

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

1 Introduction

Researchers that employ online respondents for survey experiments are often concerned about identifying and correcting for inattentive participants. Online tasks are easy to skim and respondents may not pay full attention.[1] As a result, manipulation checks are now frequently used to identify inattentive participants (Berinsky, Margolis and Sances 2016). There is no clear consensus, however, on how to measure attention or what to do with inattentive respondents. Moreover, the factual or instructional closed-ended manipulation checks that are recommended to assess attention have drawbacks of their own. Inattentive respondents may be able to guess and still pass, there is little variation between respondents when the criteria to pass is binary, and it is costly to include multiple manipulation checks of varying difficulty to distinguish attention between respondents.

I propose an alternative strategy to overcome some of these limitations that extends existing text-as-data approaches for open-ended manipulation checks. First, participants receive a text *prompt*, which includes instructions or a story, and afterward they recall what they consumed in an open-ended response. Then, I calculate the *document similarity* (Wilkerson and Casas 2017) to quantify how similar the prompt is to the participants' reply to the manipulation check. This generates a bounded, continuous, comparable measure of how attentive respondents are to the task at hand, while accounting for the content of the prompt associated with the manipulation check.

[1] Online convenience samples, such as MTurk, may include respondents that pay no less attention than other high-quality commercial or convenience samples (Thomas and Clifford 2017, 195), but "as many as half of all respondents" have displayed low levels of attention in other studies (Berinsky, Margolis and Sances 2014, 752).

Automatically computing document similarity measures allows researchers to reduce time and variation in their human-coding of open-ended manipulation checks, increase variation in attention between respondents when it exists, as well as diagnose the impact of (in)attentiveness on the results.

To examine how inattentive respondents may influence the results of mean-based comparisons, such as linear regression, I first down-weight participants by their document similarity. Specifically, I inspect how the sample average treatment effect (SATE) from a regression model using the weighted sample differs from two common approaches to estimate the population average treatment effect (PATE): (1) all participants are kept with no consideration of attention, and (2) participants that fail the manipulation check are removed from the sample.[2] The goal of comparing the weighted model to these two extremes is to distinguish if the overall effect among all participants differs from the effect among participants that likely receive the treatment, or the local average treatment effect (LATE). To highlight how inattentive participants may inhibit our ability to make inferences about the general population, I then simulate the sampling distribution of likely compliers and non-compliers.

As a proof of concept, I re-analyze an open-ended manipulation check embedded within an online survey experiment that was administered among a nationally representative sample of Americans. All elements of the re-analysis use publicly available R software that I developed to conveniently implement these guidelines for open-ended manipulation checks.

[2]Deleting respondents that fail manipulation checks forces varying levels of attrition across treatment conditions (Aronow, Baron and Pinson 2019), and any estimates of the PATE conducted on such a subsample of respondents likely discount certain demographic groups (Anduiza and Galais 2016).

2 Open-Ended Manipulation Checks as Data

I begin by describing the design requirements and assumptions associated with automatically evaluating open-ended manipulation checks. The first precondition is that respondents are presented with instructions or a story in the form of text, which I refer to as the *prompt*. This technique, therefore, is not applicable to prompts that rely on images (still or video) because there is no meaningful method of automated comparison between the prompt and the response. The method is also only applicable to experimental designs that provide respondents with some text as part of the control condition, which is recommended when the treatment is text to maintain internal validity (Steiner, Atzmüller and Su 2016). After respondents receive the prompt, they are required to briefly rephrase the content of the prompt to finish the manipulation check.[3]

Placing the open-ended response immediately after the prompt in the manipulation check is an important factor to consider. Unlike closed-ended manipulation checks, respondents do not receive double exposure or extra information that highlights a salient aspect of the prompt. Asking respondents to recall the content they consumed, therefore, should not prime or alter respondents' outcomes. As such, we can use the randomly assigned treatment of an experiment as the prompt of the manipulation check if we assume that participants are unlikely to attrite between the treatment and manipulation check, and that participants' attention is unlikely to be differentially

[3]Researchers should not permit respondents to go backward while completing the survey to minimize copying. Responses can be captured with an entry box in which participants write their reply, or participants may record an audio reply. This is a particularly useful alternative to closed-ended manipulation checks for online audio treatments, or in-person and phone interviews. I discuss the possibility of extending the current methodology to audio treatments in the Supplementary Materials.

impacted by the treatment.[4] While these are assumptions, researchers should empirically validate whether individuals exit the survey between the treatment and manipulation check, as well as whether the effect of the treatment is suppressed or magnified among participants that are likely predisposed to pay less or more attention based on the content of the treatment.

If researchers prefer to use a manipulation check that is unrelated to the treatment, they can place any prompt that is immediately coupled with an open-ended response before or after the outcome. This is akin to "treatment-irrelevant" factual closed-ended manipulation checks (Kane and Barabas 2019). Next, I create a measure of similarity between participants' responses from the manipulation check and the content that they received in the prompt.

2.2 Selecting Similarity Metrics to Measure Attention

To help overcome spelling and grammatical mistakes that may be common in typed or transcribed open-ended responses, I first re-format words to smaller segments or n-grams (Cavnar and Trenkle 1994). For instance, let n=3, which is the recommended practice for small documents (Van der Loo 2014, 120). We can represent the word "banana" by sliding a window, in our case three characters wide since n=3, across the word $\left(\boxed{\text{ban}}\,\boxed{\text{ana}}\,\boxed{\text{nan}}\,\boxed{\text{ana}}\right)$. We then count the frequency of each unique tri-gram ("ban"=1, "nan"=1, "ana" = 2). For each participant i, I store their open-ended response

[4]For instance, the first issue arises if we analyze the content of responses in audit experiments of legislators. Some legislators reply to the request and others do not, so any inference we make from the replies is biased toward the type of legislator that is already willing to engage in constituent services (Coppock 2019). Second, Montgomery, Nyhan and Torres (2018) outline an example in which researchers want to measure the impact of civics education and want to account for the willingness of participants to "comply" or receive the treatment, so a post-treatment measure of political interest is included. They state that this is not proper because political interest is itself impacted by the treatment.

(doc_{i1}) and the prompt they received (doc_{i2}) as two vectors, such that each tri-gram (n=3) in the response is $g_{1doc_{i1}}, g_{2doc_{i1}}, ..., g_{ndoc_{i1}}$ and each tri-gram from the prompt is $g_{1doc_{i2}}, g_{2doc_{i2}}, ..., g_{ndoc_{i2}}$. Then, I calculate a measure of similarity between the vectors of grams for each participant.

I start with a simple measure of similarity, the *Jaccard*, which captures the proportion of common grams (all items which are in both sets) to total grams (all items in either set) between the open-ended response and prompt.[5] One drawback of this measure is that it relies on the number of common grams, so a larger response may be judged as more similar to the prompt than a shorter response that is conceptually more alike. To avoid this, I calculate the cosine of the angle between doc_{i1} and doc_{i2}.[6] Both similarity measures are bounded from 0 to 1, such that 1 corresponds to full overlap, and 0 equals no overlap.

Since each similarity measure captures slightly different aspects of proximity and some penalize participants more harshly for minor errors, I take an average of the similarity measures for all respondents such that $1 - \left(\sum_{i=1}^{n} 1 - s_i \frac{1}{n}\right)^k$.[7] High attention respondents have a score closer to one, and the penalty k determines how severely low attention participants are down-weighted. Greater k reduces the influence of high attention respondents on the regression estimates, while increasing the potential noise from low attention respondents. I set $k = 3$ in the analysis because I want to heavily

[5] More formally, let $U(g_{doc_{i1}})$ and $U(g_{doc_{i2}})$ be the unique grams from doc_{i1} and doc_{i2}, so the intersection over the union is $s_{Jaccard(doc_{i1}, doc_{i2})} = \frac{U(g_{doc_{i1}}) \cap U(g_{doc_{i2}})}{U(g_{doc_{i1}}) \cup U(g_{doc_{i2}})}$.

[6] The cosine of the angle between the two vectors is $s_{Cosine(doc_{i1}, doc_{i2})} = \frac{U(g_{doc_{i1}}) U(g_{doc_{i2}})}{\|U(g_{doc_{i1}})\|_2 \|U(g_{doc_{i2}})\|_2}$, where $\|\cdot\|_2$ is the square root of the sum of the squared vector values $\left(\sqrt{\sum (U(g_{doc_i})^2)}\right)$.

[7] $\bar{s}_i \in [0, 1]$, hence $\lim_{k \to \infty} 1 - \bar{s}_i^k = 1$, unless a participant has perfect recall, in which case both measures will equal one and they will receive a weight of 1 regardless of the penalty (e.g., $1 = 1 - (0.5 \times (1-1) + 0.5 \times (1-1))^k$).

down-weight inattentive respondents, and respondents' similarity measures are still highly correlated with the "correct" answer as determined by a human coder.[8]

Still, n-gram similarity measures do not capture semantic meaning, which may be especially problematic for participants that articulate a clear understanding of the prompt, but select different words to describe it. Therefore, I rely on a trained word embedding technique to estimate the distance between synonyms so I can calculate how close a respondent's open-ended response is to the prompt in a semantic space (Kusner et al. 2015). Though I use n-gram measures in the manuscript because they are easy to implement and they produce comparable results to word embeddings for our example, researchers should apply the appropriate similarity measure for their context.[9] If a manipulation check references longer and more subjective text, the word embedding technique may better account for responses that are semantically equivalent to the prompt.

2.3 Diagnosing the Impact of Attention from Similarity Measures

A major problem experimentalist often face is whether they want to make a statement about how the general population selectively chooses content, or how individuals that pay attention to the experimental content react (Leeper 2017). To achieve the first goal,

[8] I explore in the Supplementary Materials how varying k alters the underlying distribution of the sample used to estimate the weighted regression models. I also discuss potential solutions if statistical power is a concern for reducing the effective number of observations in the weighted sample. Finally, given the potential biases that result from weighting, I recommend weighting by attentiveness only when the full sample is representative of the intended target population.

[9] I compare the cosine of the angle between the responses and prompts from Kane (2020) in a word embedding space, as well as in a two-dimensional space with the n-gram approach. The cosine distance similarity from the n-gram and word embeddings methods are highly related to each other ($r = 0.87$), as well as with the "correct" human answer ($r = [0.74, 0.68]$). All correlations in the manuscript and Supplementary Materials are reliable at $\alpha < 0.05$.

researchers often use weights to estimate the PATE from the SATE (Franco et al. 2017), but this is problematic if inattentive respondents are different in their mannerisms and characteristics. As such, I begin by investigating how the SATE differs when I (1) include all respondents without accounting for attention, (2) remove respondents that are assessed by a human coder to have incorrectly answered the manipulation check, and (3) down-weight participants based on their inattention. If the marginal effects fluctuate between the three models, we want to know whether it is due to inattentive respondents. Ultimately, we want to estimate the treatment of those participants that received and did not receive the treatment to understand how inattentive participants impact our ability to generalize to the larger population.

Accordingly, I simulate a sampling distribution of the average treatment effect for compliers and non-compliers varying the cutoff threshold for "receiving" the treatment. The sampling distribution of the LATE shows whether participants that engage with the treatment answer the outcome systematically different than participants that did not. First, I randomly select a cutoff from a uniform distribution bounded between zero and a user-defined threshold at the beginning of each round of the simulation. All respondents that have an average similarity measure less than or equal to that cutoff are labeled as "non-compliers". I then estimate the ATE for participants above and below the threshold to conclude one simulation round. After a sufficient number of draws, I use 100 in the manuscript, I inspect the resulting distributions of treatment effects for compliers and non-compliers. I provide guidelines on how to perform the simulations in the Supplementary Materials, including details pertaining to the selection of cutoffs and number of iterations.

Finally, researchers may be concerned that inattentiveness is associated with certain sub-groups and that if we discount inattentive individuals in our analysis we may bias our estimate of the PATE. As such, I recommend that researchers estimate a model in which attention is regressed on common socio-demographic characteristics such as age, gender, race, and education. I report in the Supplementary Materials that older, more highly educated, or white respondents are more likely to record higher levels of attention in the application from Section 3. Importantly, I do not find evidence that partisans pay more or less attention to certain treatment conditions, which would violate our central assumption that participants are not more or less attentive to treatments that they like or dislike.

If researchers wish to achieve a consistent estimator for the PATE with non-compliance, they can up-weight inattentive participants that are originally under-sampled by their average attention measure "to reflect the distribution in the target population" (Aronow and Carnegie 2013, 497). This is highly dependent, however, "on what inattentives' responses would be, were they to pay attention" because the counterfactual responses of inattentive respondents must be captured by attentive participants with similar individual characteristics to recover "the true population quantities of interest" (Alvarez et al. 2019, 158). Therefore, I suggest that practitioners are as transparent as possible and use all the information they have at their disposal to explore how different modeling decisions to address endogenous selection through attention impact the results.

3 Application: Partisan Motivated News Selection

I re-analyze the open-ended responses to the manipulation check from an existing study (Kane 2020), which explores why partisans select news stories that highlight within party consensus or disagreement in the United States. The experimental condition provides a nationally representative sample of respondents with four news stories, three of which remain constant and apolitical across respondents. The fourth news headline presents a story about President Trump, and randomizes the intra-party competition that the President faces. The first condition, the internal party *unity* treatment, states that "Trump recently pleases many of his conservative supporters", while the internal party *disunity* prompt asserts that "Trump recently upsets many of his conservative supporters". Finally, the *control* condition merely mentions "Basic biographical information about President Trump".

After respondents view this list of news stories they are asked to choose one of the news stories to read. The outcome measures whether the respondent selects the story about the President or one of the three alternative stories. Following the outcome question, the open-ended manipulation check asks respondents to briefly write what the story about President Trump concerned. Since respondents are asked the outcome question before they recall the treatment, we need to make the additional assumption that participants' recollection of the treatment during the manipulation check is the same as when they answered the outcome.

I begin by calculating each respondent's attention to the treatment using similarity measures. I show that these measures are strongly correlated with each other, as well as the "correct" answer defined by a human coder. Finally, I replicate the original findings

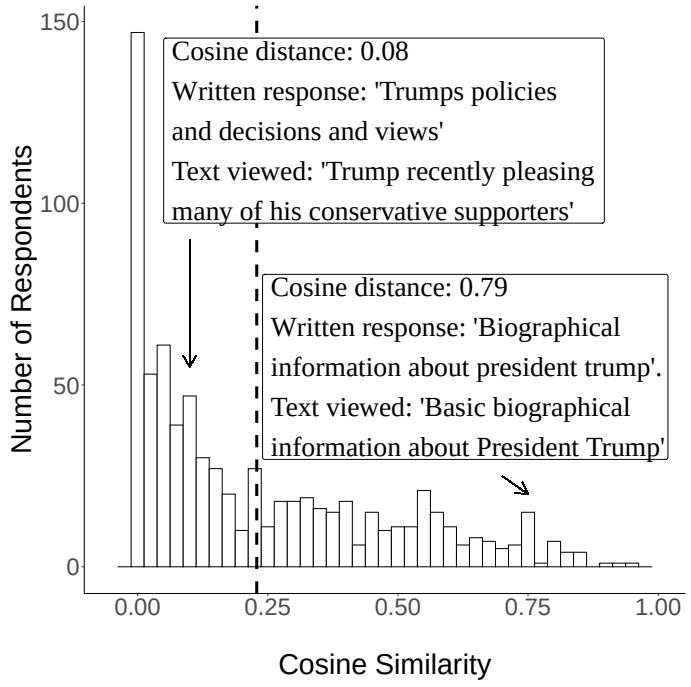
and I provide a diagnostic investigation to determine which respondents are driving the overall results. I provide all additional information regarding the re-analysis, including details on the original survey design, in the Supplementary Materials.

3.2 Measuring Attention Using Similarity Measures

Figures 1 and 2 validate that the n-gram similarity measures (1) represent the same underlying commonalities in the texts (internal validity), and (2) are related to some objective understanding of factual correctness (external validity). Figure 1 displays the distribution of cosine similarities between the treatment and participants' responses. Two open-ended responses have been selected in Figure 1 to highlight that responses closer to one are, at least subjectively to a human audience, more discernibly similar to the text they viewed as the treatment.

Figure 2 plots each respondent by their Jaccard and cosine similarity, as well as whether they correctly answered the manipulation check. First, the correlation between the two similarity measures is large, nearly one ($r = 0.98$), which signals that they capture a related latent dimension. Second, both similarity measures are strongly associated with the correct answer as determined by a human coder ($r = [0.68, 0.74]$). This lends greater credibility to the intuition that similarity measures closer to one are more likely to be factually accurate or correct. Still, since human coding can lead to large inefficiencies and inaccuracies (Kane 2020, A22), and similarity measures lack external validity without a comparison to objective correctness, it is preferable to compare both human and automated metrics of correctness when possible. Next, I use the similarity measures to see how the results differ when accounting for attention.

Figure 1: Distribution for participants' cosine similarity measures.

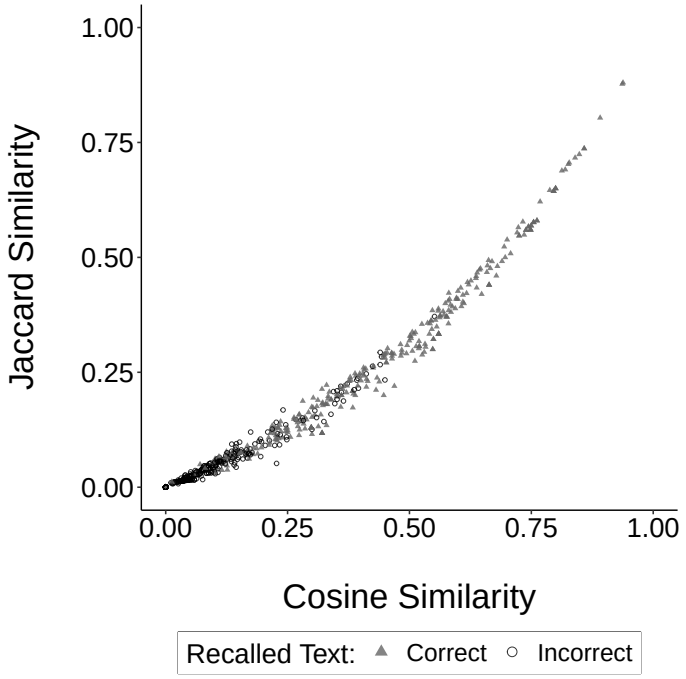


Notes: The number of participants that answered the manipulation check and the outcome in Kane (2020) is N=742. The mean cosine similarity for the sample is represented by the vertical dotted-line in Figure 1.

3.3 How Does Attentiveness Impact the Overall Results?

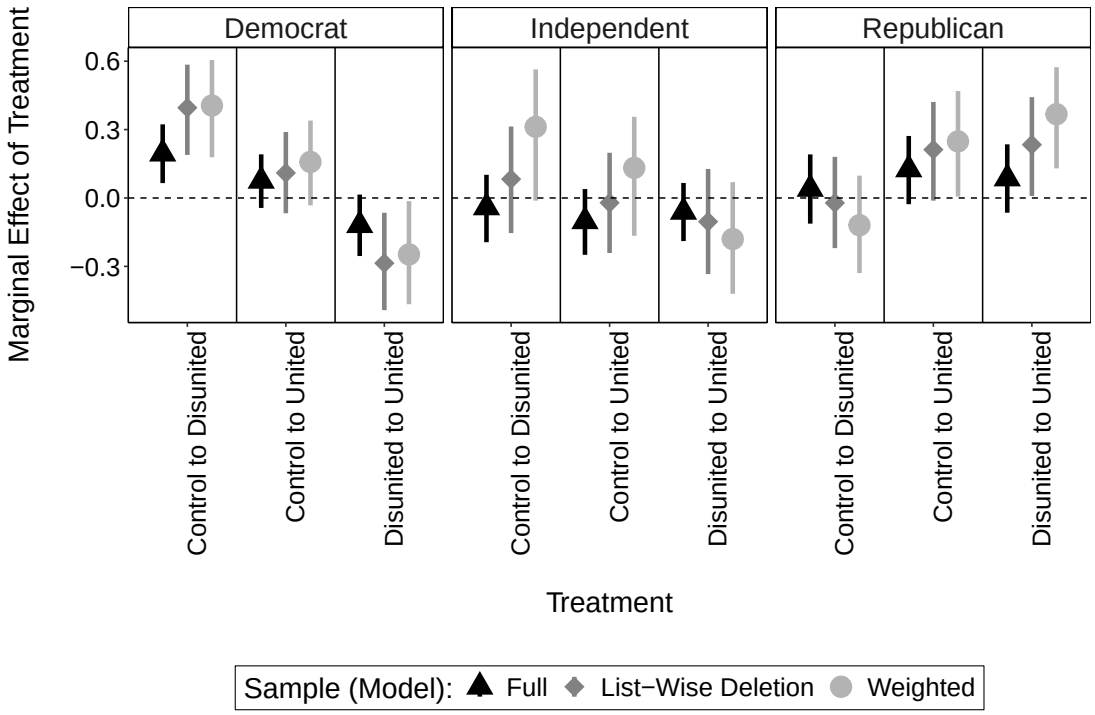
Figure 3 displays the average marginal effects of each treatment category by party identification and sample. The marginal effects estimated from the logistic regression model using the full sample irrespective of attention are shown by the black triangles. The central finding from Kane (2020) is exactly replicated: partisans are more likely to select stories about disagreement within the opposing party, but not agreement within their own party. Though the raw data suggest that Republicans are on average more likely to select the news story when it is about unity, the relationship is not statistically differentiable from zero in a regression model, which mirrors the initial findings.

Figure 2: Similarity measures by whether respondents answered manipulation check "correctly".



The second set of models, shown by the gray diamonds in Figure 3, removes respondents using list-wise deletion by whether respondents answered the manipulation check correctly, while the third set of models (represented by the gray circles) apply weights using our similarity measure. We can see in the far right panel of Figure 3, for example, that when we down-weight inattentive participants, Republicans that receive the unity prompt in comparison to the disunity prompt are more likely to select the news article about President Trump. This finding supports the first hypothesis posited by Kane, which was initially unsubstantiated in the regression that used all respondents irrespective of attention. As such, I investigate the LATE to assess whether the difference between the models is likely due to inattentive participants.

Figure 3: Marginal treatment effects by party identification, treatment category, and sample.

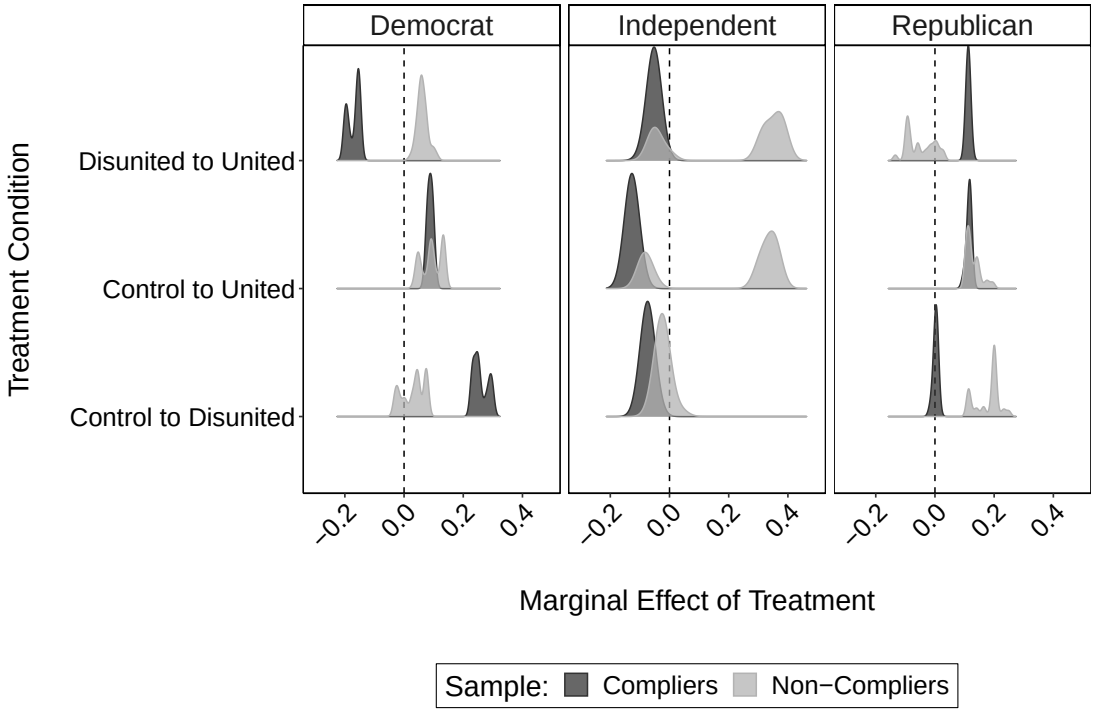


Notes: The figure plots the marginal change in the predicted likelihood that respondents select the Trump news story given a shift from the "Control" treatment (biographical content) to either a "Disunited" or "United" news story by party identification. The mean marginal effects are represented by the vertical lines. The full table of estimated coefficients from the three logistic regression models are provided in the Supplementary Materials.

3.4 Simulating the Average Marginal Effect of the "Compliers"

Figure 4 reports the distributions of average treatment effects for respondents that likely received the treatment and those that likely did not. For respondents that absorbed the treatment, which are represented by the dark gray distributions, there is little uncertainty in the ATE. The average treatment effect for respondents that did not retain the treatment, however, fluctuates widely. The wide deviations between participants that likely did not receive the treatment is further evidence that those participants are likely inattentive.

Figure 4: Distributions of average marginal treatment effects among respondents that likely received and did not receive the treatment by party identification.



Notes: The figure plots the median marginal effects of respondents that likely received and did not receive the treatment. Each distribution consists of 100 estimates of the LATE varying the threshold for "compliers".

In fact, the average effect for the compliers is distinct from the non-compliers in most of the partisan and treatment categories. Democrat compliers, found in the far left panel of Figure 4, are more likely to select a partisan news story over a neutral story, but they strongly prefer a story about a disunited Republican party instead of united. Democrat non-compliers, however, are not more or less likely to select news stories about President Trump.

Second, Republican compliers are more likely to select news stories that discuss unity

within their party. Republican non-compliers, however, are more likely to favor stories that focus on intra-party disunity. These inattentive responses pull in the opposite direction of Republican compliers, which suggests that the source of bias in the overall results comes from non-compliers. Interestingly, high attention Independents are less likely to select partisan stories in general, although inattentive Independents' responses vary greatly. This may explain the null results found in the original study.

4 Conclusion

For researchers that already utilize open-ended manipulation checks, automated similarity algorithms offer a more systematic, replicable, and transferable criteria to quantify attention than human coders. For researchers that have favored closed-ended manipulation checks in the past, similarity measures generated from open-ended manipulation checks allow for greater variation between respondents when it is present and are less likely to fall prey to participants' guessing. Additionally, I outline that researchers can use similarity measures to investigate how inattentive participants impact their ability to make inferences from their sample to the general population.

Since similarity measures can be calculated with any language that uses a written alphabet, I also introduce an application in the Supplementary Materials of an online survey experiment in Brazil and Mexico. I use this example to demonstrate how to properly construct and inspect open-ended manipulation checks with easy-to-use and open-source software that I developed in R. With these tools, analyzing open-ended manipulation checks using similarity measures is an efficient and inexpensive alternative for social scientists that rely on online respondents for surveys and experiments.

Funding

The additional applications in the Supplementary Materials were supported in part by funding from the John C. Danforth Center on Religion and Politics, as well as the Department of Political Science at Washington University in St. Louis. The study was approved by the Internal Review Board at Washington University in St. Louis (ID 201805040) in May 2018.

Acknowledgements

I very much appreciate John Kane for his support and willingness to share his data. I thank the editorial staff of Political Analysis, my anonymous reviewers, Michael Bechtel, Scott Clifford, Justin Esarey, Dino Hadzic, Jonathan Homola, Jae Hee Jung, Ryan Kennedy, Jeong Hyun Kim, Miguel Pereira, Margit Tavits, and Dalston Ward for their careful feedback, as well as participants of TADA and SPSA for constructive discussions.

Data Availability Statement

The replication materials for this paper can be found at the [link to PAN Dataverse], and the R package can be downloaded at my GitHub webpage.

Supplementary Materials

To access the supplementary materials accompanying this paper, please visit [link to PAN DOI].

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.

Anduiza, Eva and Carol Galais. 2016. "Answering Without Reading: IMCs and Strong Satisficing in Online Surveys." *International Journal of Public Opinion Research* 29(3):497–519.

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.

Aronow, Peter M., Jonathon Baron and Lauren Pinson. 2019. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* 27(4):572–589.

Berinsky, Adam J., Michele F. Margolis and Michael W. Sances. 2014. "Separating the Shirkers From the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3):739–753.

Berinsky, Adam J., Michele F. Margolis and Michael W. Sances. 2016. "Can We Turn Shirkers into Workers?" *Journal of Experimental Social Psychology* 66:20–28.

Cavnar, William B. and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175.

Coppock, Alexander. 2019. "Avoiding post-treatment bias in audit experiments." *Journal of Experimental Political Science* 6(1):1–4.

Franco, Annie, Neil Malhotra, Gabor Simonovits and L.J. Zigerell. 2017. "Developing standards for post-hoc weighting in population-based survey experiments." *Journal of Experimental Political Science* 4(2):161–172.

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.

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.

Leeper, Thomas J. 2017. "How does treatment self-selection affect inferences about political communication?" *Journal of Experimental Political Science* 4(1):21–33.

Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do About It." *American Journal of Political Science* 62(3):760–775.

Steiner, Peter, Christiane Atzmüller and Dan Su. 2016. "Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap." *Journal of Methods and Measurement in the Social Sciences* 7(2):52–94.

Thomas, Kyle A. and Scott Clifford. 2017. "Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments." *Computers in Human Behavior* 77:184–197.

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

Wilkerson, John and Andreu Casas. 2017. "Large-scale computerized text analysis in political science: Opportunities and challenges." *Annual Review of Political Science* 20:529–544.