Getting longer Responses to Open-Ended Questions in Web-Based Surveys

Abstract

Open-ended survey questions provide useful information, but their high degree of difficulty encourages respondents to satisfice and provide perfunctory answers. This experiment varied the appearance of questions asking respondents to explain the meaning of the words 'left' and 'right' in politics. Out of four tested design elements, two had an effect — asking respondents to target a (visible) word count, and a lottery scheme rewarding longer responses. No treatments seemed to reduce quality, and spending more time on longer responses only affected following closed-ended questions if they had almost identical phrasing to the open-ended ones. These interventions promise longer responses, but researchers should bear in mind the potential ethical pitfalls that come with monetary incentives and requests for longer answers.

Competing interests: The author declares none

Introduction

Open-ended survey questions provide social scientists with useful information due to the broader range, and greater depth, of responses compared to closed-ended questions. In the past, the potential benefits of this additional information had to be weighed against the increased workload of researchers tasked with coding responses into categories. Advances in text-as data methodologies over the last decade have dramatically shifted this cost-benefit calculation, allowing researchers to draw insights from datasets that include thousands of texts without armies of research assistants (Grimmer and Stewart 2013).

Researchers can estimate individuals' ideology (e.g. Lauderdale and Herzog 2016), find emergent topics (e.g. Roberts et al. 2014), or study how different groups of people use words (e.g. Rodriguez, Spirling, and Stewart 2023). A large proportion of surveys are now also conducted online, meaning researchers can be sure to capture a respondent's verbatim answer, often at a much lower cost than previously possible, rather than one paraphrased or reduced to a code by the survey enumerator.

These new techniques make open-ended survey questions more appealing to researchers, who value the additional information they provide. However, these quantitative techniques also require large amounts of text — not only do we need large sample sizes, but many techniques are aided greatly by more text per respondent (for topic models in particular, see Grimmer, Roberts, and Stewart 2022, 245).

But their unstructured nature makes open-ended questions difficult for respondents, who find the additional cognitive load burdensome. Respondents want to respond briefly, especially when they are paid for the survey and want to get through it as quickly as possible to maximize their hourly income. Those of us who commission surveys could require minimum word counts, but that virtually guarantees responses with irrelevant content, while potentially being coercive. A less coercive approach is more desirable. In this paper, I report on the results of an online experiment that instead sought to increase respondents' motivation to give in-depth responses.

Theory: Survey Response

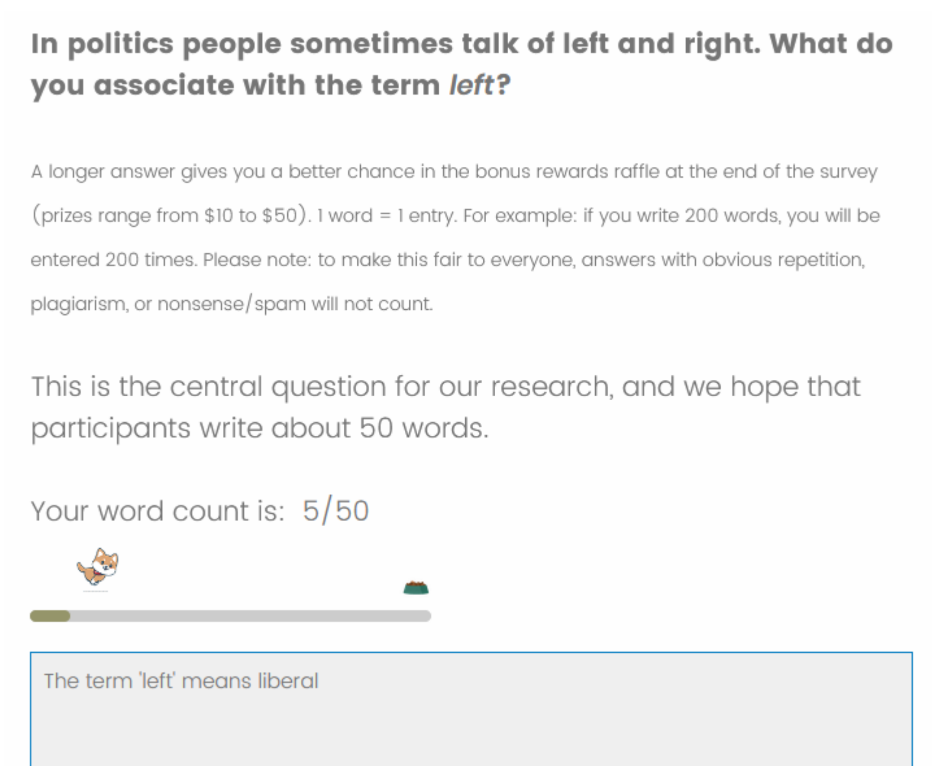
Open-ended questions can provide rich and detailed data because respondents are not constrained by a limited set of closed-ended options. Their responses, ideally, express their actual attitudes in the moment. The path to this rich response is arduous for the respondent, however. A canonical framework holds that survey respondents need to perform four tasks: they must understand what the question asks of them, search their memory for relevant information, combine this information into a single attitude, and then report this attitude to the researcher (Tourangeau 1984, 73). Absent guidelines or anchors provided by response options, every one of the cognitive tasks involved in a survey question becomes more difficult. Interpreting the question is harder without response options giving more hints as to its meaning. The search through memory for relevant considerations has to start from scratch, without anchors. Finally, translating attitudes into a response for the researchers is much harder than identifying which response option is closest to respondents' preference — they have to write prose, a task so cognitively demanding even professionals (like academics) gripe about it.

Faced with this difficulty, it is not surprising that many respondents will satisfice —– that is, devote only partial cognitive attention to each part of the process, or even skip steps entirely. Satisficing theory (Krosnick 1991, 221) holds that the likelihood of satisficing depends on three factors: the difficulty of the survey task, the ability of the respondent to answer, and the motivation of the respondent.

We are constrained in how much we can change the difficulty of survey tasks, as survey question design is primarily influenced by research needs. It is also unlikely we can change respondents' cognitive abilities. Thus, this study focuses on increasing respondents' motivation to respond.

Increasing Motivation: Treatments

Figure 1: All conditions in one treatment



We know that paying people to take surveys works: monetary incentives increase survey response rates (e.g. DeCamp and Manierre 2016). On the question level, Bullock et al. (2015) have shown that even small incentives can result in more honest answers. Because paying each respondent can be expensive, others have successfully used lottery schemes to recruit participants (Deutskens et al. 2004; Laguilles, Williams, and Saunders 2011; Zhang, Lonn, and Teasley 2017). It stands to reason that it might be possible to induce people to spend more effort on question in the same way:

- *Lottery.* In this treatment, respondents see text informing them that each word they write will provide them with additional entry into a raffle for rewards.

Visual elements of a survey have also been known to work. For example Israel (2010) shows that simply providing a larger text box for respondents induces longer responses. In the experiment, I tested two design elements:

- *Progress bar.* As respondents type more words, a bar above the text box gradually fills up and changes color, from grey to yellow to green. This mirrors a fairly common occurrence when setting a new password, where many websites and apps use a progress bar and/or changing colors to indicate the security of an entered password.
- *Animated Dog.* In a more playful variant of the progress bar, respondents see an animated image of a running dog. As they type, this dog shifts from left to right —— and closer to an illustrated bowl of food.

Finally, a third method that might motivate respondents is simply asking them to put a lot of effort into their response. It is possible to think of surveys through the lens of social exchange: some respondents might gain non-pecuniary satisfaction from feeling they have helped the researcher by providing high-quality answers (Laguilles, Williams, and Saunders 2011, 540). Indeed, researchers have obtained longer responses by simply emphasizing the importance of questions (Smyth et al. 2009). Respondents might react better to a request for more details if they had a concrete idea of what specific response length would fulfil the researcher's request:

- *Word count.* In this treatment, respondents are told that the open-ended question is central to the researchers' work, and that researchers are hoping to receive statements of a certain length (50 words). This is reinforced with a word count that updates as they type.

Data and Methods

550 respondents took a survey via the online platform CloudResearch Connect in August 2023. The survey firm implemented quotas on race, age, and gender to match proportions reported by the U.S. census. I discarded 28 either because they might be located outside the U.S. or because I suspected them of using computer-generated text in responses, leaving 522 answers.

Respondents were asked to answer two open-ended questions, asking them to define the terms *left* and *right* in the context of politics. They were also asked to respond to a few closed-ended questions on similar topics — placing themselves on the left-right scale, placing the Republican and Democratic parties on the same scale, as well as their own ideology and partisanship. There were two types of randomization: first, to study whether answering questions of one type affects responses to the other, half of the respondents began with the open-ended questions, and the others started with the closed-ended questions. Second, respondents saw a (combination of) the above treatments meant to increase their motivation to provide lengthier responses.

This study was designed as a full-factorial design, which meant that respondents might see any combination of the four treatment options (or none of them at all). Each possible combination was seen by at least 20 respondents. The study was preregistered.[1]

Results

Response Length

For simplicity, I focus on responses to the question asking respondents to describe the political left, as analyses focusing on the average of the two open-ended questions looked virtually identical.

[1] https://aspredicted.org/FT3_YMY

As a baseline, respondents who saw none of the treatments wrote about 20 words on average. Two treatments had a significant effect: the lottery and explicit request for a certain word count. Table 1 shows the average marginal effect of each treatment: respondents exposed to the word count treatment wrote 19 words more, on average, than those who were not, while those who were told about the lottery wrote around 25 words more than those who got no information about the lottery. Comparing the distribution of results in Figure 2 indicates that lottery respondents were especially likely to write unusually long responses.

Table 1: Average Marginal Effects of treatments

Treatment	Estimate	SE	p	25%	75%
color	-1.96	3.01	0.51	-7.85	3.93
dog	-1.83	3.01	0.54	-7.72	4.06
lottery	25.84	3.18	0.00	19.61	32.07
word	19.12	2.96	0.00	13.32	24.92

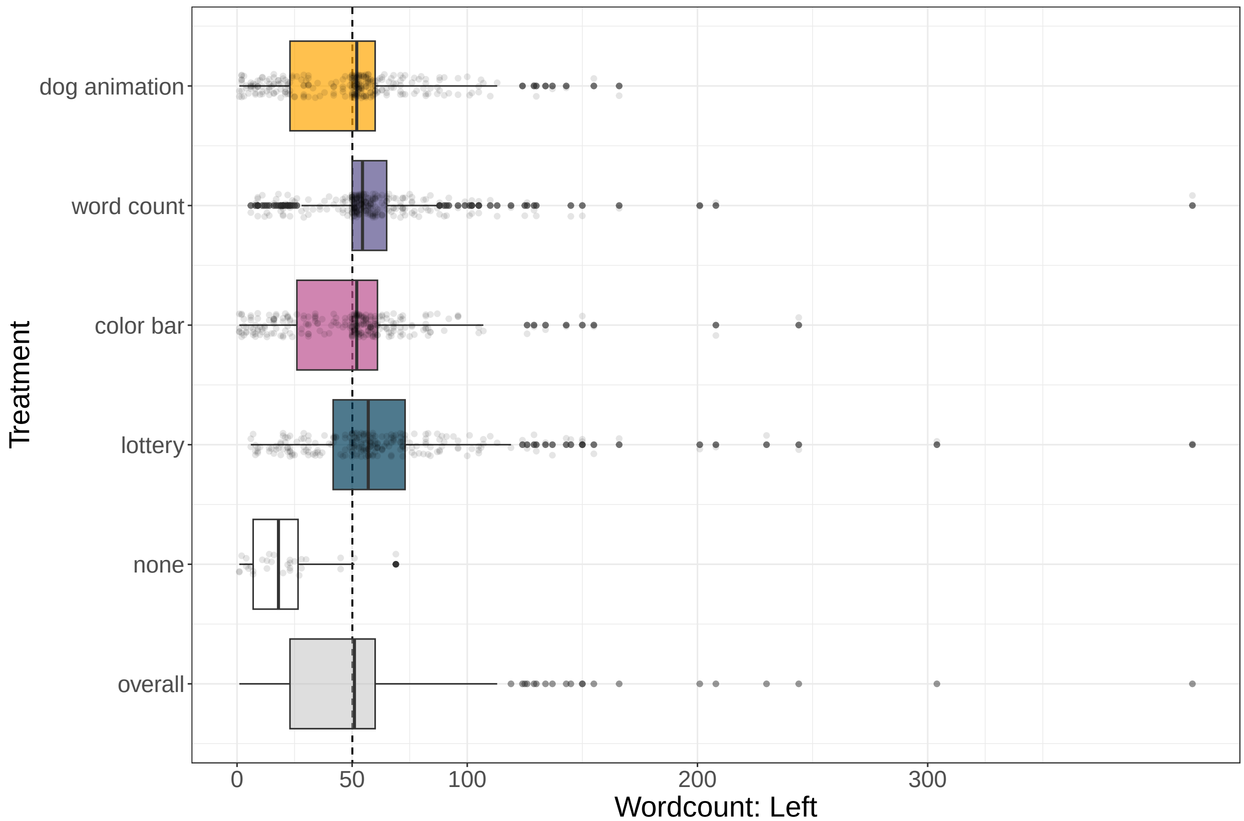
Note:

Average Marginal Effects with robust standard errors based on regression model with full interactions among all treatments. See Supplementary Materials for full table

This is a significant increase: compared to people who saw none of these treatments, the effective treatments *doubled* the amount of text respondents provided. This is a lot of effect for relatively simple and inexpensive treatments — the lottery cost just over $200 in gift cards and administration fees, and the word count treatment required only the addition of some custom code.

I used the full-factorial design primarily for its ability to test many treatments on a fairly small sample, but it also has the benefit of allowing us to test for the effect of combinations of treatments by looking at the interaction coefficients. In this case there are no major surprises: neither the dog nor the progress bar become 'activated' by contact with another treatment. It seems that the presence of the lottery treatment can compensate for the absence of the word count treatment, but

Figure 2: Distribution of wordcounts across treatments. Note that respondents saw multiple treatments at once, so most observations here show up as multiple points.



(just not quite) the other way around (see Supplemental Materials). In short, in terms of raw word count the lottery treatment slightly outshines the word count treatment, though I outline reasons for preferring the word count treatment below.

Quantity at the expense of Quality?

As pre-registered I also test whether treatments affect one of three measures of lexical diversity, as a proxy of quality: Type-Token Ratio, Mean Segmental Type-Token Ratio, or the Gini-Simpson index. The Type-Token ratio is the ratio of the count of unique words over the total number of words. The sentence: "The left is the democrats" has a ratio of $\frac{4}{5}$, because it has 5 words in total, but only 4 unique ones. The MSTTR splits the document into segments, calculates the type-token ratio for each segment, and then takes the mean across all segments. The Gini-Simpson index "represents the precise probability that any two [words] sampled randomly in succession will belong to different types" (Jarvis 2013, 93). As Table 2 shows, only Type-Token Ratio showed any significant results, with lottery and wordcount treatments leading to a reduction in the measure. This is unsurprising, as the type-token ratio virtually has to decrease with longer texts (Johansson 2008, 63; Covington and McFall 2010, 95), and those treatments cause respondents to write more.

Especially in the treatment conditions that had an effect, response quality might change over the course of the response. A respondent might begin by thoughtfully writing down all considerations that came up when they thought about the question. Then they might turn to less-relevant material, listing examples, or repeating what they had already said in order to hit the requested word count target or gain more lottery entries. I also modeled two measures of repetition in a longitudinal fashion, to see if repetition changes over the course of a response. As the Supplemental Materials show, there is no evidence that any treatment had detrimental effects by these measures.

Regardless, TTR and use of adjectives are measures of linguistic diversity or repetition — not quite the same thing as quality. To get a subjective indicator of quality, I also hand-coded results: while blinded to the treatment, I counted the number of topics in a document, and rated the quality of the response on a three-point scale. As Table 2 shows, treatments increased the number of topics, but did not affect the subjective quality.[2]

[2] In fact, some repetition in an answer might actually be useful when applying text-as data methods. In the supplementary materials, I suggest that the extra information provided by repetition might help in the case of a topic model, for example, increasing the confidence that a document belongs to a given topic.

Table 2: Quality indicators show no detrimental effects

	TTR	MSTTR	Simpson's Gini	number of topics (hand-coded)	quality rating (hand-coded)
(Intercept)	0.887 ***	0.990 ***	0.986 ***	1.333 ***	1.667 ***
	(0.020)	(0.003)	(0.002)	(0.184)	(0.098)
dog	0.007	−0.002	0.001	−0.016	0.065
	(0.026)	(0.004)	(0.002)	(0.242)	(0.130)
word	−0.148 ***	0.000	−0.002	0.719 **	−0.167
	(0.026)	(0.004)	(0.002)	(0.246)	(0.132)
lottery	−0.153 ***	0.006	−0.001	0.817 **	−0.117
	(0.032)	(0.005)	(0.003)	(0.290)	(0.156)
color	−0.025	0.001	−0.001	0.121	0.000
	(0.027)	(0.004)	(0.003)	(0.254)	(0.136)
Num.Obs.	522	501	515	507	507
R2	0.277	0.018	0.035	0.159	0.051
R2 Adj.	0.255	−0.013	0.006	0.133	0.022

+ p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001. Standard Errors in Parentheses.
Higher-order interactions omitted from table but included in the model.
Full table in Supplementary Materials.

Effects of Question Order

Motivating respondents to spend more time on open-ended responses could affect their behavior later on in the survey. When respondents write longer responses, it often means they have engaged more deeply with the question, thinking about it longer than they would have without the treatment. In this experiment, people who saw the word count and lottery treatments spent four minutes on the question on average, while those who saw none of the treatments wrote for less than two minutes. This extra time spent on considering the issue could affect their responses to closed-ended questions afterwards, having brought new considerations into their short-term memory. Thus, I briefly consider whether answering open-ended questions on a topic affects responses to closed-ended questions on the same topic.

Respondents were randomly assigned to see the open-ended or the closed-ended questions first. In terms of outcomes, I asked respondents to place themselves, as well as Republican and Democratic parties, on an 11-point left-right scale, and asked for their party identification and ideology on a standard 7-point likert scale. As Figure 3 shows, respondents who saw an open-ended response first were more likely to place themselves, as well as political parties, on the very extreme end of the range. T-tests indicate that seeing the open-ended question first increases the likelihood of picking 0 or 10 for self-placement (t = *-2.93*, $p<0.01$), or 10 for the republican party (t=*-2.31*, p = *0.02*). Results for the Democratic party were not statistically significant (t=*-1.04*, p=*0.29*). A Kolgorov-Smirnov test comparing the distributions is not significant, which is unsurprising as the distributions are virtually identical, except for the extreme categories.

The other related questions —– on party identification and ideology —– showed no apparent difference between the groups, suggesting that to the extent there is an effect, it is limited to very similar questions, and respondents do not carry this over into questions even slightly conceptually distinct. (see Supplementary Materials, Figure S4 for those graphs)

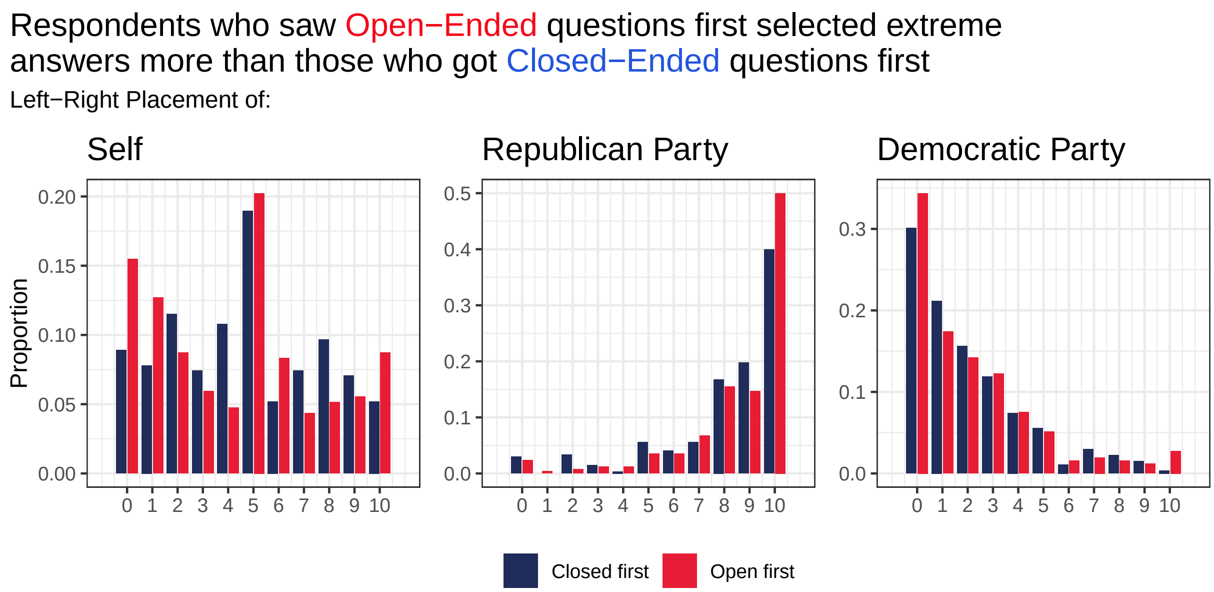
There is no effect in the opposite direction: seeing the closed-ended questions first did not affect word count, time spent on the question, or any indicators of text diversity —– see Table S11 for details. The analysis in this section is very basic, but it seems to indicate that there are few risks to getting people to spend more time on open-ended responses, as the effects on future questions are limited.

Discussion

In this paper, I report results of an experiment that aimed to increase the length of responses provided by participants in an online survey. Incentives in the form of a lottery, and explicit requests to provide more words (together with feedback on the amount of words used) proved successful, increasing average response length by 20 words or more. Feedback in the form of progress indicators did not have an effect.

Lottery and word count treatments also increased the amount of repetitiveness in answers, as measured by the type-token ratio calculated on a rolling basis. But this is no cause for alarm: human-coded quality did not suffer. Further, the effort expended on the open-ended questions

Figure 3: Effects of question order on left-right placement of oneself and political parties



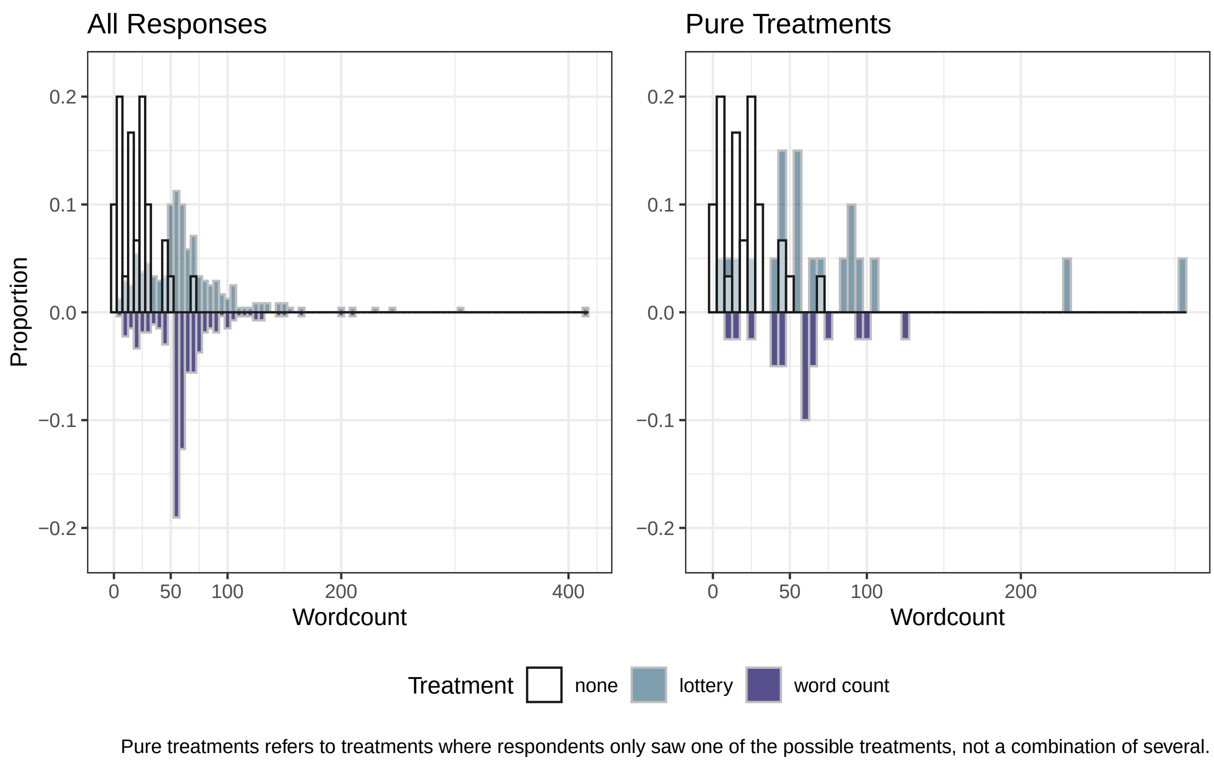
only affected answers to closed-ended questions with almost identical wording as the open-ended question.

In contrast to the lottery treatments, the word count treatment resulted in a fairly tight distribution of word counts, with a noticeable spike around 50 words (see Figure 4). It seems clear that the lottery treatment pushes respondents to dig even deeper than the word count, which gives respondents a 'finish line' at which they feel comfortable stopping.

It seems respondents might not enjoy the experience of pushing themselves to longer responses. I asked respondents at the end of the survey how the general experience of the survey was. Those in the word count condition rated the experience less positively, by 0.6 points on a 10-point scale (see Supplementary Materials Table S13).

That respondents respond similarly well to a request for more words seems encouraging. However, even a request meant to be innocuous might be interpreted differently. Respondents read "We are hoping to receive around 50 words." At least one respondent interpreted this as a requirement, messaging me on the platform to ask whether their response would be rejected due to coming in at 49 words — and therefore result in them not getting paid. On many online platforms, researchers can

Figure 4: Distribution of response length in word count condition. People in the word count condition do what's asked, often precisely



reject responses without providing a reason, and only sometimes face pushback from the platforms. Studies going forward should make explicit that the word limit is not a requirement, just an optional target.

The main concern with the lottery treatment is that monetary incentives can be coercive if they are large enough. This counts especially given some survey respondents view responding to surveys as a significant form of income. On balance, a version of the word count treatment that makes sure the target is optional might be the least objectionable.

References

Arnold, Taylor. 2017. "A Tidy Data Model for Natural Language Processing Using cleanNLP." *The R Journal* 9 (2): 1–20. https://journal.r-project.org/archive/2017/RJ-2017-035/index.html.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Bullock, John G, Alan S Gerber, Seth J Hill, and Gregory A Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10: 519–78.

Covington, Michael A., and Joe D. McFall. 2010. "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)." *Journal of Quantitative Linguistics* 17 (2): 94–100. https://doi.org/10.1080/09296171003643098.

DeCamp, Whitney, and Matthew J. Manierre. 2016. "'Money Will Solve the Problem': Testing the Effectiveness of Conditional Incentives for Online Surveys." *Survey Practice* 9 (1): 1–9. https://doi.org/10.29115/SP-2016-0003.

Deutskens, Elisabeth, Ko de Ruyter, Martin Wetzels, and Paul Oosterveld. 2004. "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study." *Marketing Letters* 15 (1): 21–36. https://doi.org/10.1023/B:MARK.0000021968.86465.00.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. "Topic Modeling in Embedding Spaces." Edited by Mark Johnson, Brian Roark, and Ani Nenkova. *Transactions of the Association for Computational Linguistics* 8: 439–53. https://doi.org/10.1162/tacl_a_00325.

Grimmer, Justin, Margaret E Roberts, and Brandon M. Stewart. 2022. *Text as Data*. Princeton; Oxford: Princeton University Press. https://press.princeton.edu/books/paperback/9780691207551/text-as-data.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF

Procedure." March 11, 2022. https://doi.org/10.48550/arXiv.2203.05794.

Israel, Glenn D. 2010. "Effects of Answer Space Size on Responses to Open-ended Questions in Mail Surveys." *Journal of Official Statistics* 26 (2): 271–85.

Jarvis, Scott. 2013. "Capturing the Diversity in Lexical Diversity." *Language Learning* 63 (s1): 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x.

Johansson, Victoria. 2008. "Lexical Diversity and Lexical Density in Speech and Writing: A Develop- Mental Perspective." *Working Papers/Lund University, Department of Linguistics and Phonetics* 53: 61–79.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5 (3): 213–36. https://doi.org/10.1002/acp.2350050305.

Laguilles, Jerold S., Elizabeth A. Williams, and Daniel B. Saunders. 2011. "Can Lottery Incentives Boost Web Survey Response Rates? Findings from Four Experiments." *Research in Higher Education* 52 (5): 537–53. https://doi.org/10.1007/s11162-010-9203-2.

Lauderdale, Benjamin E., and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (3): 374–94. https://doi.org/10.1093/pan/mpw017.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich. 2023. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments." *The Review of Economics and Statistics*, March, 1–44. https://doi.org/10.1162/rest_a_01317.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review*, January, 1–20. https://doi.org/10.1017/S0003055422001228.

Schofield, Alexandra, Laure Thompson, and David Mimno. 2017. "Quantifying the Effects of Text Duplication on Semantic Models." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, edited by Martha Palmer, Rebecca Hwa, and Sebastian Riedel, 2737–47. Copenhagen, Denmark: Association for Computational Linguistics. https:

//doi.org/10.18653/v1/D17-1290.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Mallory Mcbride. 2009. "Open-Ended Questions in Web Surveys." *Public Opinion Quarterly* 73 (2): 325–37. https://doi.org/10.1093/poq/nfp029.

Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines : Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology*, edited by Thomas B. Jabine, M. Straf, J. Tanur, and R. Tourangeau, 73–100. Washington, D.C.: National Academies Press. https://nap.nationalacademies.org/download/930.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." June 13, 2023. https://doi.org/10.48550/arXiv.2306.07899.

Zhang, Chan, Steven Lonn, and Stephanie D. Teasley. 2017. "Understanding the Impact of Lottery Incentives on Web Survey Participation and Response Quality: A Leverage-salience Theory Perspective." *Field Methods* 29 (1): 42–60. https://doi.org/10.1177/1525822X16647932.

Supplementary materials

Sample

Sample Demographics

		N	%
gender	Man	247	49.1
	Woman	256	50.9
race	Black or African American	66	13.1
	other	39	7.8
	White	398	79.1

Criteria for removing observations

The survey vendor promised a US-only sample, but the latitudes and longitudes provided by Qualtrics indicated that several respondents were outside of the US. While it is possible in theory that these respondents reside in the US and used a VPN service that located them somewhere else, I excluded them from analysis just in case.

Given the financial incentives of these online platforms — many individual tasks but with low payments per task — have seen increasing numbers of attempts to automate the process. This shouldn't matter to the results, as bots would be randomly assigned and therefore spread across treatments. Still, I wanted to focus on human subjects. Below, I report results with and without these subjects, with no substantive changes.

Open-ended questions used to be a good way to catch bots, but with the advent of large-language models that can provide text that can pass as human, things have become more complicated. A recent study estimated between 33 and 46 percent of workers on the MTurk platform used computer-generated text in a text summary task (Veselovsky, Ribeiro, and West 2023). Inspired by their practice of capturing all keystrokes, I counted *how many* keystrokes people used and compared this to the length of the texts they provided. I investigated those with extremely high differences between keystrokes and response length, and excluded those whose style was not clearly distinguishable as human.

The main models are repeated in Table S1 on the full sample (still excluding geographical outliers) and show no substantive difference to the results reported in the paper. As pre-registered, I also repeat the main model on the full sample, but excluding the first 20 responses, which had been collected at the point of pre-registration.

Table S1: Robustness check: model across different sample specifications

	original model	all respondents in US (includes suspected bots)	excluding first 20
(Intercept)	20.033**	19.267**	19.267***
	(6.194)	(5.839)	(5.796)
dog	2.036 (8.070)	3.919 (7.608)	3.919 (7.552)
word	34.692***	35.195***	34.721***
	(8.194)	(7.766)	(7.667)
lottery	57.267***	54.808***	54.808***
	(9.794)	(9.232)	(9.164)
color	3.437 (8.498)	3.604 (8.190)	5.719 (7.952)
dog × word	−4.386	−6.413	−5.846
	(11.396)	(10.821)	(10.619)
dog × lottery	−23.229+	−18.417	−21.262+
	(12.798)	(12.180)	(11.975)
word × lottery	−38.433**	−37.883**	−36.958**
	(12.591)	(12.018)	(11.782)
dog × color	−2.764	0.636	−2.891
	(11.483)	(10.998)	(10.708)
word × color	−5.977	−4.606	−6.558
	(11.984)	(11.585)	(11.214)
lottery × color	−28.019*	−23.976+	−26.400*
	(12.874)	(12.255)	(12.007)
dog × word × lottery	15.903	12.399	14.459
	(17.211)	(16.400)	(16.001)
dog × word × color	8.197	3.235	6.507
	(16.350)	(15.648)	(15.204)
dog × lottery × color	35.363*	31.306+	37.987*
	(17.391)	(16.537)	(16.163)
word × lottery × color	22.806	20.729	21.272
	(17.553)	(16.899)	(16.396)
dog × word × lottery × color	−33.616	−31.803	−36.744
	(24.133)	(22.981)	(22.376)
Num.Obs.	522	514	533
R2	0.230	0.243	0.242
R2 Adj.	0.207	0.221	0.220

+ p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001
full interactions included in model but omitted from table

Wordcount

Regression Tables: Wordcount

I pre-registered using a model without interactions. In the paper I use average marginal effects based on the model with all interactions, given its superior properties (Muralidharan, Romero, and Wüthrich 2023), but the substantive differences between models are very slim.

Table S2: Model without interactions

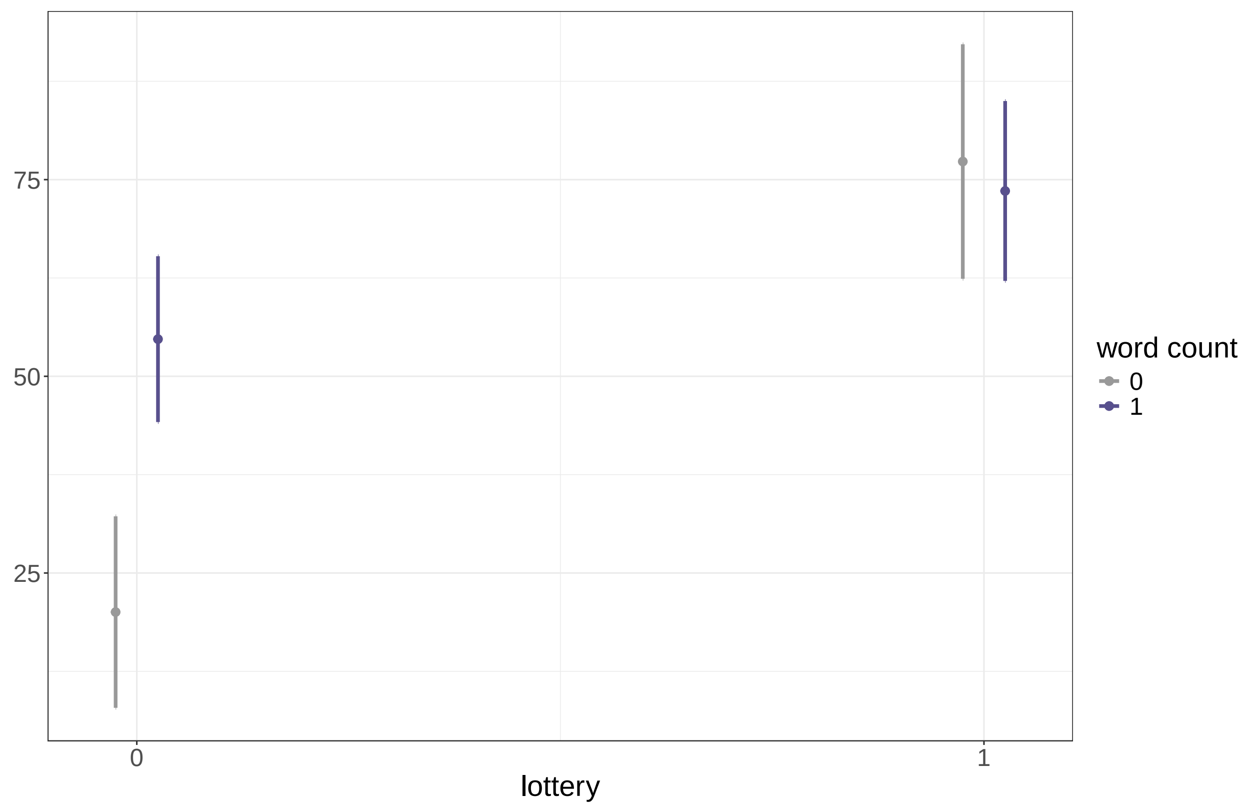
	No Interactions	Full Interactions
(Intercept)	29.545 (3.413)***	20.033 (6.194)**
dog	−1.601 (3.026)	2.036 (8.070)
word	19.617 (3.028)***	34.692 (8.194)***
lottery	25.575 (3.034)***	57.267 (9.794)***
color	−1.112 (3.024)	3.437 (8.498)
dog × word		−4.386 (11.396)
dog × lottery		−23.229 (12.798)+
word × lottery		−38.433 (12.591)**
dog × color		−2.764 (11.483)
word × color		−5.977 (11.984)
lottery × color		−28.019 (12.874)*
dog × word × lottery		15.903 (17.211)
dog × word × color		8.197 (16.350)
dog × lottery × color		35.363 (17.391)*
word × lottery × color		22.806 (17.553)
dog × word × lottery × color		−33.616 (24.133)
Num.Obs.	522	522
R2	0.186	0.230
R2 Adj.	0.180	0.207

+ p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

Interactions between treatments

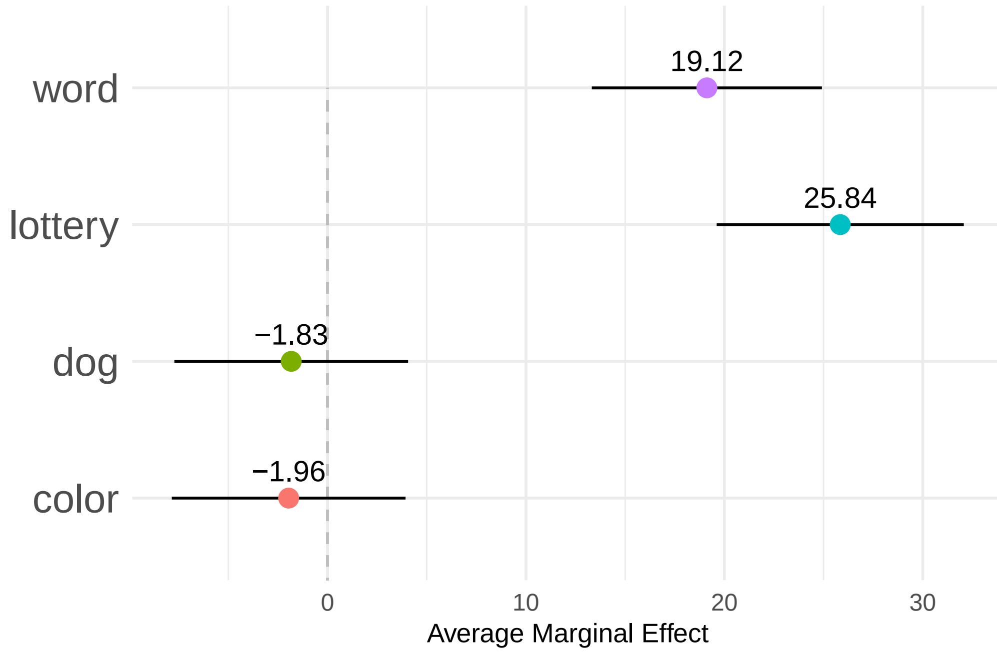
The plots for the predicted value from the interactions of word count and lottery are depicted in Figure S1. They suggests that the lottery treatment is slightly more potent than the word count treatment: A breakdown of word counts for every possible treatment combination suggests the same, see Table S3).

Figure S1: Lottery versus wordcount treatments



AME graph

Figure S2: Average Marginal Effects of Treatments



Mean Wordcounts by treatment

Table S3: Mean wordcount for every treatment combination

dog	word	lottery	color	wordcount_left
0	0	1	0	77.30
0	1	1	0	73.56
0	1	1	1	65.81
1	0	1	1	64.12
1	1	1	0	63.88
1	1	1	1	63.31
1	0	1	0	56.11
1	1	0	1	55.27
0	1	0	0	54.73
0	0	1	1	52.72
1	1	0	0	52.38
0	1	0	1	52.19
0	0	0	1	23.47
1	0	0	1	22.74
1	0	0	0	22.07
0	0	0	0	20.03

Lexical Diversity and Hand-Coded Quality Measures

Table S4: Lexical diversity and hand-coded quality measures regressed on treatments

	TTR	MSTTR	Simpson's Gini	number of topics (hand-coded)	quality rating (hand-coded)
(Intercept)	0.887	0.990	0.986	1.333	1.667
	(0.020)***	(0.003)***	(0.002)***	(0.184)***	(0.098)***
dog	0.007	−0.002	0.001	−0.016	0.065
	(0.026)	(0.004)	(0.002)	(0.242)	(0.130)
word	−0.148	0.000	−0.002	0.719	−0.167
	(0.026)***	(0.004)	(0.002)	(0.246)**	(0.132)
lottery	−0.153	0.006	−0.001	0.817	−0.117
	(0.032)***	(0.005)	(0.003)	(0.290)**	(0.156)
color	−0.025	0.001	−0.001	0.121	0.000
	(0.027)	(0.004)	(0.003)	(0.254)	(0.136)
dog × word	0.026	0.005	0.002	0.447	0.080
	(0.037)	(0.005)	(0.003)	(0.343)	(0.184)
dog × lottery	0.025	−0.001	0.000	−0.018	−0.077
	(0.041)	(0.006)	(0.004)	(0.384)	(0.206)
word × lottery	0.127	−0.005	0.000	−0.294	0.041
	(0.041)**	(0.006)	(0.004)	(0.376)	(0.202)
dog × color	0.033	0.006	0.005	−0.115	−0.232
	(0.037)	(0.006)	(0.003)	(0.345)	(0.185)
word × color	0.032	−0.001	0.001	−0.248	−0.056
	(0.039)	(0.006)	(0.004)	(0.358)	(0.192)
lottery × color	0.069	−0.008	0.000	−0.146	−0.050
	(0.042)+	(0.006)	(0.004)	(0.383)	(0.205)
dog × word × lottery	−0.037	−0.003	−0.001	−1.051	−0.212
	(0.056)	(0.008)	(0.005)	(0.519)*	(0.278)
dog × word × color	−0.052	−0.009	−0.005	0.050	0.130
	(0.053)	(0.008)	(0.005)	(0.490)	(0.263)
dog × lottery × color	−0.136	−0.001	−0.007	0.444	0.388
	(0.056)*	(0.008)	(0.005)	(0.522)	(0.280)
word × lottery × color	−0.054	0.006	0.003	0.164	0.081
	(0.057)	(0.008)	(0.005)	(0.524)	(0.281)
dog × word × lottery × color	0.136	0.004	0.004	−0.065	−0.186
	(0.078)+	(0.011)	(0.007)	(0.725)	(0.389)
Num.Obs.	522	501	515	507	507
R2	0.277	0.018	0.035	0.159	0.051
R2 Adj.	0.255	−0.013	0.006	0.133	0.022

Higher-order interactions omitted from table but included in the model. Full table in Appendix

Longitudinal Models

It is possible that repetitiveness is dynamic: respondents initially write with little repetition, until they have exhausted, and then turn to rephrasing their previous statements to extend the length of their response.

I create two rolling measures of repetitiveness: the cumulative number of adjectives used,[3] and the type-token ratio, the ratio of unique terms (types) to overall words (tokens) in the response up until that point. Both of these statistics are measured at every word in a given response, allowing me to treat the data as longitudinal data, where each additional word is one step further in the 'time' variable.

If we simply used the cumulative adjectives or Type-Token Ratio for the whole response as an outcome, both would no doubt be significant: of course the total number of adjectives goes up as people write more, and of course the Type-Token Ratio decreases with more length, as people are forced to re-use words like articles and conjunctions. Table S5 bears this out, with significant coefficients for the `wordcount` variable, which tracks the passage of time in this longitudinal model.

The advantage of a longitudinal model is that we can tell whether these measures of repetitiveness change over the course of an answer. Perhaps adjectives increase modestly, and then take a sharp upturn close to 50 words as people try to hit the target. I include a spline in the regression at the mean word count of those who have not seen any treatments (20 words). The idea is that this is the 'natural' amount of words respondents might write, if not motivated by a treatment to write more. After this point, we are more likely to see an uptick in repetition as respondents work hard to hit their desired word count. In fact, type-token ratios *increase* after the spline, pointing to a more varied use of language.

Table S5: Do treatments encourage repetition?

[3] In retrospect, this is probably an imperfect measure of repetitiveness. People with a larger vocabulary might use more adjectives, for example; more detailed description is not the same as repetition. Because this analysis was pre-registered, I report it here. Adjectives were counted using the R package `cleanNLP` (Arnold (2017)).

	cumulative adjectives	rolling TTR
(Intercept)	0.034(0.158)	1.010***(0.010)
wordcount	0.125***(0.014)	−0.008***(0.001)
lottery	0.004(0.236)	−0.028+(0.014)
word	0.011(0.203)	−0.021+(0.012)
color	0.096(0.217)	0.005(0.013)
dog	0.055(0.205)	−0.005(0.012)
spline	−0.064(0.109)	0.014*(0.006)
wordcount × lottery	−0.017(0.019)	0.005***(0.001)
wordcount × word	−0.014(0.017)	0.003***(0.001)
wordcount × color	−0.006(0.019)	0.001(0.001)
wordcount × dog	0.014(0.018)	0.002**(0.001)
SD (Intercept doc_id)	0.795	0.048
SD (wordcount doc_id)	0.060	0.003
Cor (Intercept~wordcount doc_id)	−0.256	−0.241
SD (Observations)	0.690	0.041
Num.Obs.	31 636	31 630

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Higher-order interactions omitted from table but included in the model. Full table in Appendix

Table S6: Longitudinal models

	cumulative adjectives - unconditional	cumulative adjectives	rolling TTR - unconditional	rolling TTR
(Intercept)	−0.106 (0.036)**	0.034 (0.158)	0.991 (0.002)***	1.010 (0.010)***
wordcount	0.121 (0.003)***	0.125 (0.014)***	−0.005 (0.000)***	−0.008 (0.001)***
lottery		0.004 (0.236)		−0.028 (0.014)+
word		0.011 (0.203)		−0.021 (0.012)+
color		0.096 (0.217)		0.005 (0.013)
dog		0.055 (0.205)		−0.005 (0.012)
spline		−0.064 (0.109)		0.014 (0.006)*
wordcount × lottery		−0.017 (0.019)		0.005 (0.001)***
wordcount × word		−0.014 (0.017)		0.003 (0.001)***
lottery × word		−0.397 (0.301)		0.024 (0.018)
wordcount × color		−0.006 (0.019)		0.001 (0.001)
lottery × color		−0.386 (0.313)		0.005 (0.019)
word × color		−0.334 (0.293)		−0.013 (0.018)
wordcount × dog		0.014 (0.018)		0.002 (0.001)**
lottery × dog		−0.242 (0.309)		0.001 (0.019)
word × dog		−0.295 (0.279)		0.016 (0.017)
color × dog		−0.130 (0.290)		−0.003 (0.017)
lottery × spline		0.152 (0.125)		−0.091 (0.007)***
word × spline		0.312 (0.121)**		−0.034 (0.007)***
color × spline		0.439 (0.147)**		−0.033 (0.009)***
dog × spline		0.311 (0.142)*		−0.014 (0.008)+
wordcount × lottery × word		0.008 (0.024)		−0.004 (0.001)***
wordcount × lottery × color		0.033 (0.026)		−0.002 (0.001)*
wordcount × word × color		0.003 (0.024)		−0.001 (0.001)
lottery × word × color		0.866 (0.419)*		0.004 (0.025)
wordcount × lottery × dog		0.004 (0.025)		−0.002 (0.001)+
wordcount × word × dog		0.013 (0.023)		−0.002 (0.001)+
lottery × word × dog		0.719 (0.411)+		−0.009 (0.025)
wordcount × color × dog		−0.002 (0.026)		0.000 (0.001)
lottery × color × dog		0.329 (0.422)		−0.024 (0.025)
word × color × dog		0.465 (0.397)		0.011 (0.024)
lottery × word × spline		−0.392 (0.144)**		0.039 (0.008)***
lottery × color × spline		−0.513 (0.169)**		0.075 (0.010)***
word × color × spline		−0.481 (0.167)**		0.035 (0.010)***
lottery × dog × spline		−0.034 (0.164)		0.025 (0.010)**
word × dog × spline		−0.666 (0.161)***		0.008 (0.009)
color × dog × spline		−0.646 (0.195)***		0.024 (0.011)*
wordcount × lottery × word × color		−0.026 (0.033)		0.004 (0.002)*
wordcount × lottery × word × dog		−0.026 (0.032)		0.002 (0.001)
wordcount × lottery × color × dog		−0.036 (0.035)		0.001 (0.002)
wordcount × word × color × dog		−0.021 (0.033)		0.001 (0.002)
lottery × word × color × dog		−1.179 (0.574)*		−0.002 (0.035)
lottery × word × color × spline		0.926 (0.200)***		−0.065 (0.012)***
lottery × word × dog × spline		0.663 (0.194)***		0.005 (0.011)
lottery × color × dog × spline		0.012 (0.225)		−0.060 (0.013)***
word × color × dog × spline		0.967 (0.224)***		−0.038 (0.013)**
wordcount × lottery × word × color × dog		0.070 (0.045)		−0.001 (0.002)
lottery × word × color × dog × spline		−1.168 (0.270)***		0.056 (0.016)***
SD (Intercept doc_id)	0.801	0.795	0.054	0.048
SD (wordcount doc_id)	0.060	0.060	0.003	0.003
Cor (Intercept~wordcount doc_id)	−0.240	−0.256	−0.332	−0.241
SD (Observations)	0.692	0.690	0.043	0.041
Num.Obs.	31 636	31 636	31 630	31 630
R2 Marg.	0.690	0.649	0.657	0.571
R2 Cond.	0.989	0.988	0.978	0.975
ICC	1.0	1.0	0.9	0.9

Hand-Coding

Average Marginal Effects — Hand-coded Measures

Table S7: Average Marginal Effects on number of topics (hand-coded), with robust standard errors.

Treatment	Estimate	SE	p	25%	75%
color	0.01	0.09	0.87	-0.16	0.19
dog	0.00	0.09	0.97	-0.17	0.18
lottery	0.46	0.09	0.00	0.28	0.64
word	0.50	0.09	0.00	0.32	0.67

Table S8: Average Marginal Effects on response quality (hand-coded), with robust standard errors.

Treatment	Estimate	SE	p	25%	75%
color	-0.05	0.05	0.29	-0.14	0.04
dog	0.01	0.05	0.89	-0.09	0.10
lottery	-0.12	0.05	0.01	-0.21	-0.03
word	-0.15	0.05	0.00	-0.24	-0.06

Structural Topic Model

The usefulness of repetition for topic modeling

Take the topic model, a workhorse of quantitative text analysis in the social sciences, which assigns texts a probability of containing any one of a number of topics, each of which in turn are defined by the words that they tend to feature (e.g. Blei 2012).

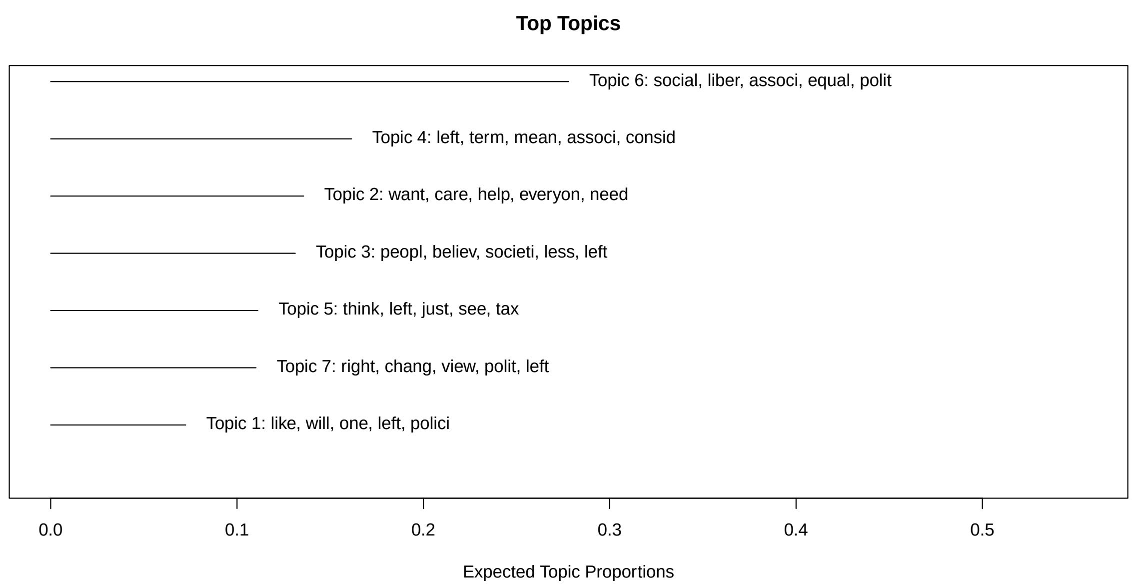
I ran a very simple structural topic model (Roberts et al. 2014), which yielded seven topics, of which one was associated with words around 'care', reflecting the many responses that mention the political left striving to care for disadvantaged groups.

One respondent whose response was assigned to this topic wrote: "They care about others. They want everyone to be equal. Take care of the sick. The poor. Give3 (sic) everyone a leg up when they need it. Take care of Seniors."

It stands to reason that, across the corpus, many of the words in that sentence tend to coexist in discussions of the same topic, and that the topic modeling algorithm will pick up on this co-occurrence. Additional words on the same topic are more information the model can use to determine that this topic is being used.[4]. Indeed, I show below that if we remove some of the extra detail — references to caring for seniors, the poor, and sick — the model is 20% less certain that this document belongs in the topic about caring for others. (see Table S10).

Take a simple, but realistic topic modeling workflow. After calculating models across a variety of values of K (number of topics), I used exclusivity and semantic coherence to pick K = 7. The topics are as follows.

Figure S3: Overview: Structural Topic Model



[4] Repetition can be a problem if entire documents are repeated (Schofield, Thompson, and Mimno 2017) or the same string appears throughout the text, but that is quite a different scenario from the one described above. I suspect the helpfulness of some repetition will only be strengthened by the inclusion of word embeddings into the topic modeling workflow (see e.g. Dieng, Ruiz, and Blei (2020) or Grootendorst (2022)), because words with similar meanings will be near each other in the embedding space.

The topic model returns, for each document, a number indicating the probability of it belonging to each of the seven topics. These numbers are called thetas. A document high on theta 1 is likely to contain writing that fits into topic 1, and so on.

For some circumstantial evidence, I calculated each document's highest theta, yielding the probability that it belongs in whatever topic the model thinks is its most likely topic. The documents with the highest probabilities — where the model is most certain about them belonging to a certain topic — often resulted from respondents seeing the word count condition. Viewed from this angle the word count treatment does not seem harmful for the purposes of text analysis.

Table S9: Top thetas by treatment

highest theta	word	lottery	dog	color
0.80	1	0	1	1
0.79	1	0	1	1
0.78	1	0	1	0
0.78	1	1	1	0
0.78	1	1	0	0
0.77	1	1	0	1
0.74	1	0	1	0
0.74	0	0	0	0
0.74	1	0	0	0
0.73	0	0	0	0

A sample text illustrates the potential benefits of repetition in text analysis. I chose this text because it scored highly on the theta for topic 2, a particularly coherent topic. The model assigned this document a high probability (0.77) of belonging to this topic. Removing some of the repetition around caring reduces the theta by 0.14, or 18% of the original theta's size.

Table S10: Topic prediction with and without repetition

document	theta	text
original (id=236)	0.77	They care about others. They want everyone to be equal. Take care of the sick. The poor. Give3 everyone a leg up when they need it. Take care of Seniors. The left still has these values but they too fall into thye gutter at times. They give in too quickly. They need to continue to fight for everyone.
amended	0.63	They care about others. They want everyone to be equal. The left still has these values but they too fall into thye gutter at times. They give in too quickly. They need to continue to fight for everyone.

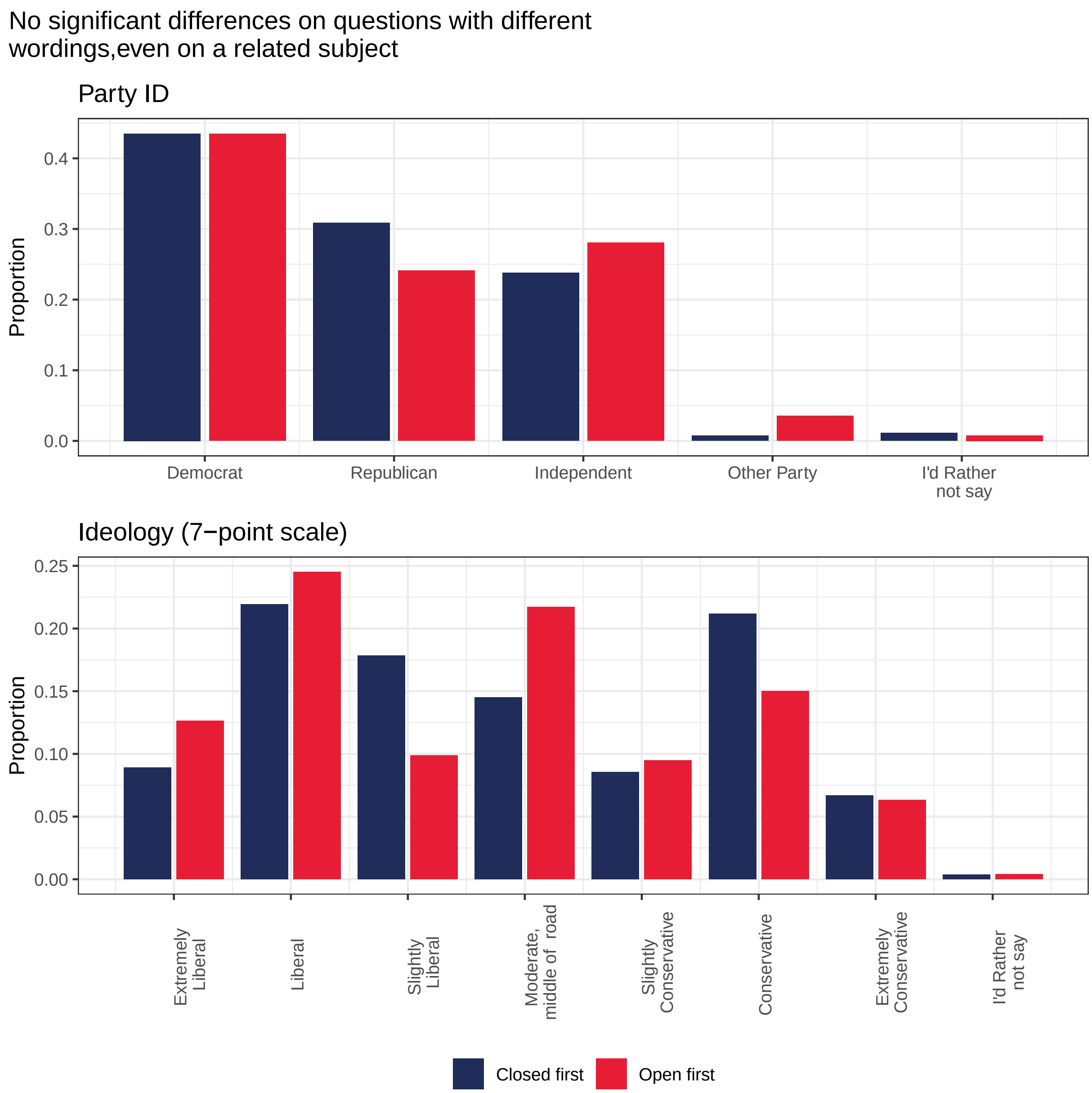
Effects of viewing Closed-ended Questions First

Table S11: Effects of viewing closed-ended questions first on measures of repetition in open-ended question

	wordcount	TTR	MSTTR	D	time spent
(Intercept)	103.565 (4.312)***	0.785 (0.008)***	0.991 (0.001)***	0.014 (0.001)***	212.941 (11.216)***
closed_first	−6.974 (6.006)	−0.004 (0.011)	0.001 (0.001)	0.001 (0.001)	−16.523 (15.624)
Num.Obs.	522	522	501	515	522
R2	0.003	0.000	0.000	0.001	0.002
R2 Adj.	0.001	−0.002	−0.002	−0.001	0.000

Effects of viewing Closed-ended Questions First

Figure S4: Effects of viewing closed-ended questions first on other questions



Effects on participant ratings

The below regression models show the effect of the treatments on the ratings participants gave for the survey's compensation, time, user experience, fairness, and overall ratings. Participants were able to give these ratings on the platform after taking the survey; as such this is a self-selected sample. I also asked one question on general experience at the end of the survey — this is the last model shown (experience (in-survey)). Interactions between all treatments are included in the model but not shown for ease of presentation.

Table S12: Effect of treatments on participant ratings

	overall	fairness	experience	time	compensation	experience (in-survey)
(Intercept)	4.929 (0.120)***	5.000 (0.115)***	4.875 (0.126)***	5.000 (0.114)***	5.000 (0.198)***	8.767 (0.355)***
dog	−0.192 (0.158)	0.000 (0.162)	0.125 (0.178)	0.000 (0.161)	0.000 (0.290)	−0.511 (0.463)
word	−0.373 (0.160)*	0.000 (0.175)	0.125 (0.193)	0.000 (0.174)	−0.429 (0.290)	−1.292 (0.470)**
lottery	0.000 (0.169)	0.000 (0.162)	0.125 (0.178)	0.000 (0.161)	0.000 (0.280)	−0.917 (0.561)
color	−0.095 (0.160)	−0.154 (0.146)	−0.029 (0.160)	−0.077 (0.145)	−0.077 (0.251)	−0.149 (0.487)
Num.Obs.	298	179	179	179	178	521
R2	0.071	0.077	0.066	0.125	0.098	0.047
R2 Adj.	0.021	−0.007	−0.020	0.044	0.015	0.018

Table S13 presents Average Marginal Effects for the outcome on general satisfaction with the survey experience as rated in-survey.

Table S13: Average Marginal Effects of treatments

Treatment	Estimate	SE	p	25%	75%
color	-0.09	0.17	0.59	-0.42	0.24
dog	0.14	0.17	0.41	-0.19	0.47
lottery	0.04	0.17	0.80	-0.29	0.38
word	-0.60	0.17	0.00	-0.93	-0.27

Note:

Average Marginal Effects with robust standard errors based on regression model with full interactions among all treatments.

Sample Texts

The below table lists, for each condition, the response whose average word count (across left and right) comes closest to the average word count for that treatment. To select these responses, I filtered the dataset to only those responses that saw only one of the treatments (or none at all). I then calculated the mean word count in each of these 'pure' treatments, and selected the respondent whose average wordcount was closest.

Table S14: Example responses

Information	Open-ended (left)

Treatment:lottery Mean wordcount (treatment):77.3 Mean wordcount (respondent): 72	|When I think of the term left I think of someone who is liberal thinking and liberal minded. The word progressive tends to come up but I think that is not true for the left. Progressive means making progress so when the left says they are progressive what it means is that we want everyone to think like we do. I think that it is best to say the left is liberal
Treatment:word Mean wordcount (treatment):54.73 Mean wordcount (respondent): 54	|The left like crazy. They love a lot of pronouns and like to yell racist if someone disagrees with them. They love power and the more government they can put in your life to control you the better for them. They love criminals, illegals and the homeless. They put them first over most Americans.
Treatment:color Mean wordcount (treatment):23.47 Mean wordcount (respondent): 22	|Left is woke and woke is loving, caring, helpful, etc. In other words, being a good Boy Scout, I'm there for it.
Treatment:dog Mean wordcount (treatment):22.07 Mean wordcount (respondent): 22	|I associate democratic views with the left. More government involvement and social programs. I think of Universal Healthcare, climate change, government programs.

Treatment:control Mean wordcount (treatment):20.03 Mean wordcount (respondent): 2	|For me, "left" is associated with high taxes, permissive attitudes towards crime and criminals, and a lax policy regarding immigration.

Code

The JavaScript code to give feedback on wordcount, as well as the other treatments, can be accessed here

Survey Questions

Table S15: survey questions

left_right_self. In politics people sometimes talk of left and right. Where would you place yourself on this scale? [Multiple Choice, 0-10]

left_right_rep. Using the same scale, where would you place the Republican Party?[Multiple Choice, 0-10]

left_right_dem. Using the same scale, where would you place the Democratic Party?[Multiple Choice, 0-10]

party_US. Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what? [Multiple Choice, Democrat; Republican; Independent; Other Party (text entry option); I'd rather not say]

ideology. When it comes to politics, how would you describe yourself? [Multiple Choice: Extremely Liberal; Liberal; Slightly Liberal; Moderate, middle of the road; Slightly Conservative; Conservative; Extremely conservative; I'd rather not say]

open_l_n. In politics people sometimes talk of left and right. What do you associate with the term *left?* [open-ended, might include other text or visual design elements as per treatment condition described in main article] *open_r_n.* In politics people sometimes talk of left and right. What do you associate with the term *right?* [open-ended, might include other text or visual design elements as per treatment condition described in main article]

general. Finally, on a scale of 1 to 10, where 0 is very unpleasant and 10 is very pleasant, how do you rate your experience taking this survey? [Multiple Choice, 0-10]