Supporting Information

Details of how models were selected for inclusion in the specification curve analysis

We consider 4 categories of variation in model specification for inclusion in the specification curve: decisions about (a) which covariates to include in the model, (b) which set of study days to include in the model, (c) whether to subset data by state, and (d) whether to apply inverse probability weights to the data.

Because, in our judgment, a number of the analytical choices the replicating authors (1) made are not reasonable, we conduct the specification curve analysis in two stages. We begin with a specification curve that includes an exhaustive list of model specifications that we consider to be reasonable. This first-stage specification curve is not limited to specifications we believe to be optimal but, rather, includes all specifications we believe a competent expert with a reasonable understanding of the psychological theory underlying the original paper might use. We then expand the set of specifications to include ones that reflect most of the analytical decisions the replicating authors made (see Table S1).

Randomization check. To inform our judgment about what model specifications are reasonable and warranted, we began with an analysis to determine whether there were any chance imbalances in assignment to experimental conditions; that is, whether any of the key baseline characteristics of participants that could plausibly affect voter turnout (i.e., the covariates listed above) were unequally distributed across experimental conditions despite random assignment. In this analysis, we addressed two technical issues with the randomization check reported by Gerber and colleagues (1) in their paper (see below for detail). Because randomization to condition was implemented separately by the two survey firms that collected data, we regressed experimental condition on each of the demographic variables in turn, always also including an indicator variable for survey firm and for the interaction between survey firm and the relevant demographic variable. Because the purpose of this analysis was not to draw generalizable inferences from systematic patterns in the data but rather to identify chance imbalances that are large enough that they could plausibly influence the treatment effect estimate, we use a threshold of $p < 0.1$ (two-tailed) to classify a chance imbalance as significant enough to warrant including a model in the specification curve that adjusts for it statistically using covariates. Our analysis revealed imbalances across conditions for each of the following baseline characteristics that were significant at $p < 0.1$ (two-tailed) in data from either SSI or YouGov, and/or an interaction with survey firm that was significant at $p < 0.1$ (two-tailed): state (imbalance for all 4 states), race (Black), and turnout history (the 2004, 2010, 2012, and 2014 general elections and the 2006 and 2008 primary elections). (See below for additional detail.)

Covariates. Gerber and colleagues (1) included 7 types of covariates in all of their models: survey firm, gender, race, state, date on which data were collected, number of days before the election on which data were collected, and vote history.

Survey firm. Data for this experiment were collected by two private research firms, which implemented random assignment independently: YouGov and Survey Sampling International

(SSI). Both firms use opt-in sampling methods (as opposed to probability sampling) so it is plausible that the populations each firm sampled differ and that there might be other differences in how the firms interact with panel members. Moreover, it is widely seen as necessary to adjust for any variable on which random assignment is stratified (2, 3). Gerber and colleagues (1) included an indicator for survey firm in all of the models they reported, and we include such an indicator in all specifications included in the specification-curve. That is, we do not treat this as a decision about which there is potential disagreement since both we and the replicating authors appear to agree that including an indicator for survey firm is necessary.

Gender. Gender is a basic demographic variable that investigators frequently adjust for and that is known to influence turnout (4). Gerber and colleagues (1) adjusted for gender in all of their models.

Race. Race is a basic demographic variable that investigators frequently adjust for in models and that is known to influence turnout (5). Gerber and colleagues (1) adjusted for race in all of their models.

State. Voter turnout often varies substantially by state (as it does in the present data) and, in the present experiment, state is also a proxy for substantially different election contexts, so it seems reasonable to adjust for state. Gerber and colleagues (1) included indicators for state in all of their models.

Date of data collection. While the elections in Kentucky, Mississippi, and Houston were all held on Tuesday, November 3rd, 2015, the election in Louisiana was held on Saturday, November 21st, 2015. Gerber and colleagues (1) adjusted for date in all of their models.

Number of days before the election. The present experiment was conducted over 4 days, beginning three days before the election and ending when polls closed on Election Day. The replicating investigators included a covariate for the number of days before the election on which data were collected (range: 0 to 3). We note that treating number of days as a quantitative variable (rather than using a separate indicator variable for each possible number of days before the election) imposes the assumption that any effect of number of days before the election is linear, which does not seem to us to be a safe assumption.

Vote history. Having voted in past elections is a strong predictor of voting in future elections (6), so controlling for it seems reasonable. In their models, Gerber and colleagues (1) included indicator variables for turnout in 15 different elections, all of which were held in the 11 years before the study.*

* They controlled for turnout in the general elections in 2004, 2006, 2008, 2010, 2012, and 2014, the primary elections in 2004, 2006, 2008, 2010, 2012, 2014, and 2015, and "presidential primary" elections in 2008 and 2012. It is not clear to us how the "presidential primary" indicator variables differ from the "primary election" variables for 2008 and 2012. Moreover, 4 of the turnout history variables included in the replicating authors' models (the 2012 primary, the 2012 presidential primary, the 2014 primary, and the 2015 primary) lack data from at least one state, introducing additional sources of potential error into their model. Therefore those 4 turnout history variables were excluded from consideration as reasonable covariates.

Sub-setting data by day. The original experiments by Bryan and colleagues (7) were conducted only the day before and early (before 9 am) the morning of election day. This design choice was based on our presumption that any motivation triggered by the noun wording was likely to be relatively fleeting and therefore could only be expected to translate into increased turnout if the treatment were administered very close to the election. For example, it is possible that a fleeting boost in motivation to vote is only translated into actual voter turnout if that motivation prompts a person to immediately make an implementation plan for how they will carry out their intention to vote (8). This seems most likely to occur if one experiences that boost in motivation very close to Election Day. Bryan and colleagues' decision to end data collection early the morning of Election Day was similarly deliberate. It was based on the logic that a boost in motivation to vote is less likely to translate into actual voting if one does not have much time to get to the polls before they close. As Election Day progresses, it becomes less and less likely that people will be able to find time to go to the polls before they close. For these reasons, models that include only data collected the day before the election provide the closest approximation of the original experiment. Models including data collected two and three days before the election are also reasonable but should be understood to be testing extensions of the original theory that probe how long before an election the noun treatment might still be effective at boosting turnout. Models including data from Election Day are not a reasonable test of the treatment effect. If data were available that included the time of day at which participants were treated, it would be reasonable to include data from early the morning of Election Day but the publicly available data do not contain this information. Gerber and colleagues' (1) main analysis is a specification including data from all days. They also report specifications that include data from one day before and the day of the election (combined—they do not report any test using only data from the day before Election Day).

Sub-setting data by state. In addition to their primary models, which include data from all 4 states in which the experiment was conducted, the replicating authors report models including only data from Kentucky and Louisiana (1). Their rationale for sub-setting by state in this way is that the gubernatorial elections in each of those states was that they were both rated "toss up" races by the Cook Political Report before the election and were therefore unambiguously competitive. This test was framed as a response to the original authors' criticism of the first replication test, which was conducted in overwhelmingly uncompetitive primaries for the 2014 midterm elections—elections that failed to produce a context in which the identity "voter" was likely to feel important and worthwhile (9). The point of the original authors' critique, however, was not that an election's competitiveness is the primary determinant of whether it provides the necessary psychological context. Rather, the point was that primaries for midterm elections are, by default, of such low salience that most registered voters are likely not even to be aware of them. One exception to that general rule would be if a primary were both competitive and meaningful—that is, if the outcome of the primary were in doubt and the winner of the primary could plausibly win the subsequent general election.[†] So, the original authors' emphasis on competitiveness was specific to the midterm primary context in which the first replication test was conducted. We can see no clear basis on which to argue that any subset of

[†] For example, in jurisdictions that overwhelmingly favor one party, that party's primary election is typically the "main event"—the election that essentially determines who will be elected. By contrast, the opposite party's primary election is generally regarded as inconsequential since the nominee is extremely unlikely to be elected.

the four general elections the present study was conducted in would be more or less likely to create the necessary psychological context for the noun-vs.-verb effect to manifest. For example, even if competitiveness were the correct criterion, the 2015 Houston mayoral election was decided by a smaller margin than either of the two gubernatorial elections the replicating authors singled out as highly competitive. Presumably that race was not rated a "toss up" by the Cook Political Report because that publication did not provide ratings of races for local offices.[‡] The turnout rate among participants in the verb condition, which seems a more reasonable gauge of interest in and attention to each election was highest in Houston (51.5%) and lowest in Louisiana (34.4%). The Mississippi election, which was by far the least competitive of the four (winning margin: 34.1 percentage points) was a close second to Houston in terms of turnout in the verb condition (49.7%). In sum, we can see no defensible argument for sub-setting the data by state. Given that sub-setting data comes at a big cost in terms of statistical power, which biases results toward failures to replicate, we do not believe this is a reasonable analytical decision.

Inverse probability weights. Because the replicating authors do not provide an explanation of how the weights were computed or what purpose they are meant to serve, we are unable to evaluate the reasonableness of applying them.

Stage 1: Reasonable model specifications. Next, considering the chance imbalances in participants' baseline characteristics revealed by our randomization check, we compiled a list of reasonable analytical decisions about which we believe competent experts could disagree. While including data from 2 or 3 days before the election is not reasonable as a direct replication of the original experiments, including those data as tests of an extension of the original experiments is reasonable. Including data from Election Day is unambiguously not reasonable for the reasons articulated above. We also did not include specifications that subset the data by state because doing so has no benefit (e.g., in terms of clarity of interpretation) and imposes a major cost (in terms of statistical power). We also did not include specifications that apply inverse probability weights because the replicating authors provide no information about how they were computed or what they are for. Finally, including 15 separate indicator variables for turnout in previous elections and their higher-order interactions with state, resulting in a large number of highly collinear covariates, is not reasonable. Four of the turnout history variables included in the replicating authors' models (the 2012 primary, the 2012 presidential primary, the 2014 primary, and the 2015 primary) lack data from at least one state, introducing additional potential sources of error and/or bias into their model. Thus, we determined that a reasonable solution would be to include, as covariates, indicator variables for some but not all of the previous elections in the data set. Because we identified chance imbalances in condition assignment on the indicators for turnout in 6 of those elections (the 2004, 2010, 2012, and 2014 general elections and the 2006 and 2008 primary elections) in data from at least one of the survey firms, we selected those as the covariates one could reasonably include in a model.

This left us with a list of 9 analytical decisions about which we believe competent experts could reasonably disagree. They include (a) whether or not to include gender as a covariate, (b) whether or not to include race as a covariate, (c) whether or not to control for the interaction

[‡] The ballot for the Houston mayoral election also included 7 statewide ballot measures proposing amendments to the Texas state constitution, many of which attracted substantial attention.

Table S1. Complete list of variations in model specification included in the specification curve analysis. Unshaded cells indicate variations included in both the Stage-1 and Stage-2 specification curves. Shaded cells indicate variations included only in the Stage-2 specification curve. The complete set of specifications included in each curve was determined by fully crossing all variants then deleting any models that were perfectly redundant with others already in the model.

Class of Decisions	Decision Elements	Number of Variants	Variants
Covariate specification	Gender: *indicator for male, indicator for unknown*	2	(1) No gender covariates; (2) include indicators for male and unknown gender as covariates
	Race: *indicator for Black, indicator for Hispanic, indicator for other*	3	(1) No race covariates; (2) include indicators for Black, Hispanic, and other as covariates; (3) include indicators for Black, Hispanic, and other and interaction between Black and survey firm
	State: *indicator for LA, indicator for MS, indicator for TX*	3	(1) No state covariates; (2) include indicators for LA, MS, and TX as covariates; (3) include indicators for LA, MS, and TX and interaction between each of those and survey firm as covariates
	Number of days before Election Day when participants were treated: *indicator for treatment 1 day before Election Day, indicator for treatment 2 days before Election Day, indicator for treatment 3 days before Election Day*	2	(1) No covariates for number of days before Election Day; (2) include indicators for treatment 1 day before Election Day, treatment 2 days before Election Day, and treatment 3 days before Election Day
	Vote History: *indicator for turnout in 2004 general election, indicator for turnout in 2006 primary election, indicator for turnout in 2008 primary election, indicator for turnout in 2010 general election, indicator for turnout in 2012 general election, indicator for turnout in 2014 general election*	3	(1) No vote history covariates; (2) include all 6 indicators for vote history as covariates; (3) include all 6 indicators for vote history and interactions between each and survey firm as covariates
	Exact set of 95 covariates included in Gerber and colleagues' (2018) analyses (see The Present Analysis section for complete list)	1	(1) Include all 95 covariates
Subsample	Participants treated on Election Day, participants treated 1 day before Election Day, participants treated 2 days before Election Day, participants treated 3 days before Election Day	3	(1) Include participants treated 1 day before Election Day only; (2) include participants treated 1 or 2 days before Election Day; (3) include participants treated 1, 2, or 3 days before Election Day
		3	(1) Include participants treated on Election Day and 1 day before; (2) include participants treated on Election Day and 1 or 2 days before; (3) include participants treated on Election Day and 1, 2, or 3 days before
Weighting	Apply inverse probability weights, no weights	2	(1) Apply inverse probability weights (2) Do not weight data

between the Black racial category and survey firm to correct for the chance imbalance across conditions on that variable in the YouGov sample, (d) whether or not to control for state, (e) whether or not to control for the interaction between state and survey firm to correct for the chance imbalance across conditions on that variable in both survey firms (in opposing directions), (f) whether or not to control for the number of days before the election on which data were collected, (g) whether or not to control for turnout in the 6 elections listed above, (h) whether or not to control for the interaction between survey firm and the 6 elections listed above to control for chance imbalances across conditions on those variables, and (i) whether to include data from 1, 2, and 3 days before the election, only from 1 and 2 days before, or only from the day before the election. These decisions were then fully crossed with each other, resulting in a set of 324 different model specifications. Of those, 54 were completely redundant with other specifications included in the curve and so were deleted (e.g., models controlling for number of days before the election are completely redundant with models that do not control for this in the subset of data from a single day). The final Stage-1 specification curve included 270 different models (see Table S1).

Stage 2: Omnibus set of model specifications. Next, we compiled a larger set of possible specifications that includes decisions we believe are unreasonable but that we know the replicating authors believe to be reasonable at least insofar as they were included in the specifications those authors reported in their paper (1). This omnibus set includes all models included in the Stage-1 specification curve plus a model with the exact set of covariates the replicating authors included in their analyses. It also adds specifications using subsets of data that include Election Day (i.e., the day before and day of the election, the 2 days before and day of the election, the 3 days before and day of the election) and models that apply inverse probability weights. The result is a set that includes 1,308 different model specifications. Of those, 108 were completely redundant with other specifications included in the set and so were deleted. The final Stage-2 specification curve included 1,200 different models (see Table S1).

Details of the randomization check in the present analysis and problems with Gerber, Huber, and Fang's (1) **randomization check.**

When we examined Gerber and colleagues' publicly posted analysis syntax, we identified two technical issues with the randomization check reported in their paper (1). First, Gerber and colleagues did not control for the interaction between covariates and survey firm. Their sample was collected by two separate professional firms—SSI and YouGov—and those firms each randomly assigned their participants to experimental conditions independently. Failing to include the interaction between potential covariates and survey firm can cause the analysis to miss cases in which imbalances differ in data from the two firms. This is particularly problematic because Gerber and colleagues did not control for interactions with survey firm in their primary model specifications testing for treatment effects (1). In fact, there were a number of instances of imbalances across conditions that differed across the two firms, and failing to include the interaction of the relevant baseline covariate with survey firm would result in a failure to control for those imbalances, possibly biasing results.

Second, when the replicating authors tested for imbalances across conditions using categorical variables, they tested interactions with only one group constituting the excluded category, but did

not rotate the excluded category, so one level of each categorical variable was excluded from the randomization check. Since the choice of the excluded category in the dummy variable coding is arbitrary, it is preferable to test all pairwise combinations by rotating the excluded category.

We tested all potential covariates (i.e., variables that could plausibly predict turnout in the target elections), using a series of simple linear regressions (without survey weights), to determine whether they were successfully balanced between the treatment and control groups in data from each survey firm. First, we estimated separate regressions, one for each potential baseline covariate, and included their interaction with a dummy variable indicating the survey firm that was coded such that YouGov was the contrast category (0 = YouGov, 1 = SSI). The baseline covariates were state, day, gender, race/ethnicity, and all variables indicating voting in prior elections in the dataset (see complete list above). As noted, all categorical variables were tested separately with dummies that rotate the contrast category (e.g., first KY, MS, and LA were compared to TX, then LA, KY, and TX were compared to MS, etc.). For each regression, we noted whether a given baseline covariate predicted treatment status at $p < 0.10$ (which would indicate a substantial imbalance across conditions within the YouGov sample), and we noted whether there was a significant interaction between the covariate and firm at $p < 0.10$, which would indicate that the balance across conditions was different in data from the two firms. Next, we re-estimated the same regressions, re-coding the indicator variable for firm so that SSI was the contrast category (0 = YouGov and 1 = SSI). This allowed us to test for imbalances within the SSI sample.

Details of multicollinearity in Gerber and colleagues main model specification

The variance inflation factor (VIF) of a predictor in a linear model is the standard measure of multicollinearity. Common recommended cutoffs, above which a VIF value is considered potentially problematic are 4, 5, and 10. All of those values are very high and there is broad agreement that a VIF above 10 indicates extreme collinearity (10–13). For example, using the present data and regressing turnout on 10 predictors that one would correctly presume would have a substantial degree of multicollinearity with each other (survey firm, 2 indicators for gender, 3 indicators for race, and an indicator for prior voter turnout in the most recent previous general election), VIF values range 1.03 to 1.90 and the mean VIF across all predictors in the model is 1.25. In the replicating authors' main analysis, 54 covariates have a VIF greater than 4, 43 of those have a VIF greater than 5, 7 of those have a VIF greater than 10, and 2 of those have a VIF greater than 90. The mean VIF across all predictors in their model is 6.77.

Stage-1 specification curve results among participants treated 1 or 2 days before Election Day

The Stage-1 specification curve, including data only from participants who completed the manipulation 1 or 2 days before the election, contained 108 different models, 100% of which yielded a result with a one-tailed *p*-value less than 0.05 ($p_{\text{specification curve}} < 0.0005$, by the statistical significance metric). The median effect size point estimate in this specification curve was 4.2 percentage points ($p_{\text{specification curve}} = 0.008$, by the effect size metric), the same as the corresponding Stage-2 specification curve.

Using Bayesian Causal Forest (BCF) to test for heterogeneity in the treatment effect

To test the possibility that administering the treatment 1 or 2 vs. 3 days before the election is a true source of systematic heterogeneity in the noun-vs.-verb treatment effect, we implemented the "Bayesian Causal Forest" (BCF) algorithm (14).

BCF is a flexible Bayesian model that is designed to uncover true sources of heterogeneous effects while requiring few, if any, decisions from the researcher, thus minimizing opportunities to exercise researcher degrees of freedom (14). Since BCF incorporates a strong prior presumption that effect sizes are centered at zero and that groups do not differ from each other (or that, if they do, the differences are small), the results are conservative. BCF goes beyond recent advances in "Bayesian Additive Regression Trees" (BART) (15), which have been used prominently in field experiments in political science (16). BCF has outperformed BART in public, head-to-head competitions in which the objective was to maximize the accurate detection of systematic treatment effect heterogeneity while minimizing the risk of false positives (14, 17). So, BCF is considered one of the most promising methods currently available for detecting heterogeneity in treatment effects, while requiring little researcher intervention and minimizing the risk of false positives.

The BCF method has several advantages over the traditional, hands-on, linear, frequentist regression approach. First, it can discover the best-fitting model specification, including non-linearities and higher-order interactions, using a machine learning, "sum of trees" approach. This means that researchers do not have to make arbitrary choices about which covariates to include, or which interactions among them to include, but instead can rely on the algorithm's search of the data to provide a better fit. Second, BCF is designed to reduce confounding in the estimation of treatment effects. It does so in part by separating out a function to explain the main effects of variables from the function to explain the interactions between moderators and the treatment. This is important in the present experiment because many variables that predict voting were not evenly distributed across the treatment and control groups, despite random assignment. Simulation studies (14) show that moderation tests can be biased in experiments where random assignment has failed to produce even distributions of potential moderators across conditions, leading researchers to falsely attribute moderation to confounders, or to fail to attribute moderation to variables that are confounded with the treatment. Thus, BCF can provide more accurate moderation tests.

We applied the BCF algorithm to the replication study data collected 1, 2, and 3 days before the election. We did not include data collected on Election Day, because those data do not provide a valid test of the hypothesis due to design degrees of freedom exercised by the replicating authors and discussed in the main text. In addition to study day, we provided BCF with data on survey firm as well as all potential covariates that we identified as having substantial chance imbalances across conditions: gender: male; race: Black; state: KY, TX, MS, and LA; and vote history: 2004 general, 2010 general, 2012 general, 2014 general, 2006 primary, 2008 primary, and 2012 primary.

The BCF algorithm yielded two main conclusions. First, affirming the conclusion from the specification curve analysis, BCF found that noun wording increased voter turnout when the

noun-vs.-verb manipulation was administered 1 or 2 days before the election. An examination of the posterior density distribution, which ranges from 0 to 6 percentage points with 95% confidence, showed a very high posterior probability (19:1 odds) that the effect among participants treated those days was greater than zero, ($\overline{CATE}_{1,2\ days\ before}$ = 2.8 percentage points, $pr(\overline{CATE}_{1,2\ days\ before} > 0)$ = .95. Note that a typical frequentist *p*-value for a group difference is not the same as the posterior probability that two groups differ, so this finding should not be misinterpreted as the equivalent of a *p*-value of 0.05 (for explanations of how p-values are commonly confused with Bayesian probabilities, see Refs 18–20). Moreover, BCF employs a conservative prior distribution that shrinks the posterior distribution (and therefore estimated effects) toward zero (i.e., no effect). This also applies to tests for heterogeneity—BCF conservatively shrinks posterior distributions toward homogeneity—which further shrinks simple effects toward zero when data include subgroups with no effect. Therefore, this result is much stronger evidence than a *p*-value of 0.05, since this posterior probability was updated by the data from a strong prior presumed probability of .5 (i.e., 1:1 odds, or no effect). That conservative shrinkage toward zero also applies to BCF's point estimates so BCF's point estimate of the effect among participants treated 1 or 2 days pre-election (2.8 percentage points) should be treated as conservative. In sum, the data from this replication experiment provide sufficiently strong evidence to warrant a very large shift in posterior probability toward the conclusion that noun wording has a positive effect on voter turnout when administered either one or two days before Election Day. Even the conservative point estimate of the size of the (intent-to-treat) effect of noun wording is as large as the complier average causal effect (CACE) estimate of the effect of the more costly GOTV method of calling individual households.

Second, affirming another conclusion suggested by the frequentist specification curve analysis, the BCF algorithm found that the noun-vs.-verb manipulation was very unlikely to have a positive effect on voter turnout when administered 3 days before the election ($\overline{CATE}_{3\ days\ before}$ = -.0004 percentage points, 95% posterior density interval -4 to 4 percentage points), posterior $pr(\overline{CATE}_{3\ days\ before} > 0)$ = .51. Thus, the data provided little evidence to change the prior probability of .50 that the group mean for participants who completed the manipulation survey on Day 3 was greater than zero.

Finally, the BCF algorithm found strong evidence in the data to warrant the conclusion that the effect among participants treated 1 or 2 days before Election Day was greater than the effect among those treated 3 days before Election Day, $pr(\overline{CATE}_{1,2\ days\ before} > \overline{CATE}_{3\ days\ before})$ = .96 (24:1 odds) (see Fig S1 for boxplots depicting the posterior distributions), with an expected average difference of 3 percentage points between the two subgroups. This shift in probabilities from a prior presumed probability of .5 indicates that the data warrant the strong conclusion that, in this sample, there is systematic heterogeneity in the noun-vs.-verb treatment effect such that it is present among participants treated 1 or 2 days before Election Day but is 3 percentage points smaller (and unlikely to be greater than zero) among participants treated 3 days prior to the election. Because this is the first study to detect this moderation pattern, this should still be considered only preliminary evidence for a more general phenomenon until it is confirmed by additional studies in other election contexts.

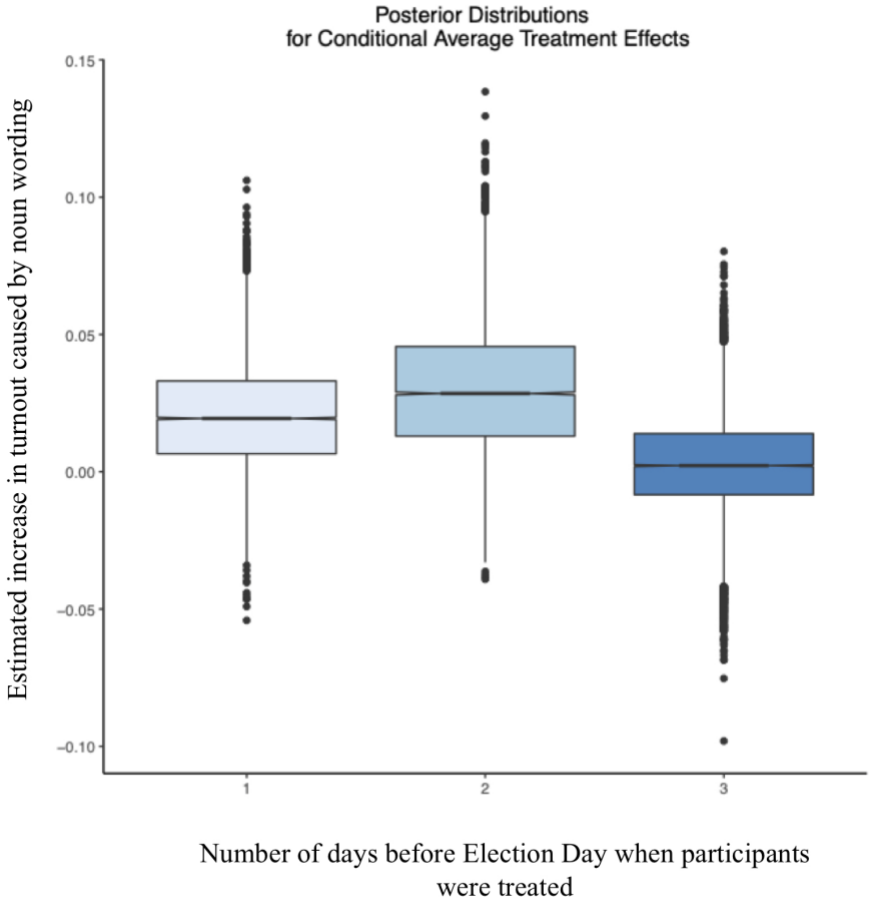

Figure S1. A Bayesian Causal Forest analysis shows that the noun-vs.-verb treatment effect on voter turnout is likely to be positive and greater than zero when participants complete the experimental manipulation 1-2 days before the election, but not 3 days before. Dots correspond to random draws from the posterior distributions of the conditional average treatment effects.

Supporting References

1. A. Gerber, G. Huber, A. Fang, Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout? Evidence From a New Replication Experiment: Outsized Turnout Effects of Subtle Linguistic Cues. *Political Psychology* **39**, 925–938 (2018).

2. A. S. Gerber, D. P. Green, *Field Experiments: Design, Analysis, and Interpretation*, 1st edition (W. W. Norton & Company, 2012).

3. R. Glennerster, K. Takavarasha, *Running Randomized Evaluations: A Practical Guide* (Princeton University Press, 2013).

4. H. Coffé, C. Bolzendahl, Same Game, Different Rules? Gender Differences in Political Participation. *Sex Roles* **62**, 318–333 (2010).

5. M. McDonald, Voter Turnout Demographics - United States Elections Project. *United States Elections Project* (November 30, 2018).

6. K. Arceneaux, D. W. Nickerson, Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments. *American Journal of Political Science* **53**, 1–16 (2009).

7. C. J. Bryan, G. M. Walton, T. Rogers, C. S. Dweck, Motivating voter turnout by invoking the self. *PNAS* **108**, 12653–12656 (2011).

8. P. M. Gollwitzer, Implementation Intentions. *American Psychologist*, 11 (1999).

9. C. J. Bryan, G. M. Walton, C. S. Dweck, Psychologically authentic versus inauthentic replication attempts. *Proc Natl Acad Sci USA* **113**, E6548 (2016).

10. J. Cohen, P. Cohen, S. G. West, L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed* (Lawrence Erlbaum Associates Publishers, 2003).

11. R. M. O'Brien, A Caution Regarding Rules of Thumb for Variance Inflation Factors in (2007).

12. M. H. Kutner, J. Neter, C. J. Nachtsheim, W. Li, *Applied Linear Statistical Models w/Student CD-ROM*, 5th International edition (McGraw-Hill Education, 2004).

13. K. P. Vatcheva, M. Lee, J. B. McCormick, M. H. Rahbar, Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)* **6** (2016).

14. P. R. Hahn, J. S. Murray, C. Carvalho, Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv:1706.09523 [stat]* (2017) (November 30, 2018).

15. J. L. Hill, Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* **20**, 217–240 (2011).

16. D. P. Green, H. L. Kern, Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly* **76**, 491–511 (2012).

17. V. Dorie, J. Hill, U. Shalit, M. Scott, D. Cervone, Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv:1707.02641 [stat]* (2017) (November 30, 2018).

18. J. Cohen, The earth is round (p < .05). *American Psychologist* **49**, 997 (19950401).

19. D. H. Krantz, The Null Hypothesis Testing Controversy in Psychology. *Journal of the American Statistical Association* **94**, 1372–1381 (1999).

20. S. Greenland, C. Poole, Living with P Values: Resurrecting a Bayesian Perspective on Frequentist Statistics. *Epidemiology* **24**, 62 (2013).