Replicator degrees of freedom allow publication of misleading failures to replicate

Christopher J. Bryan[a,1], David S. Yeager[b], and Joseph M. O'Brien[b]

[a]Booth School of Business, University of Chicago, Chicago, IL 60637; and [b]Department of Psychology, University of Texas at Austin, Austin, TX 78712

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved October 22, 2019 (received for review June 28, 2019)

In recent years, the field of psychology has begun to conduct replication tests on a large scale. Here, we show that "replicator degrees of freedom" make it far too easy to obtain and publish false-negative replication results, even while appearing to adhere to strict methodological standards. Specifically, using data from an ongoing debate, we show that commonly exercised flexibility at the experimental design and data analysis stages of replication testing can make it appear that a finding was not replicated when, in fact, it was. The debate that we focus on is representative, on key dimensions, of a large number of other replication tests in psychology that have been published in recent years, suggesting that the lessons of this analysis may be far reaching. The problems with current practice in replication science that we uncover here are particularly worrisome because they are not adequately addressed by the field's standard remedies, including preregistration. Implications for how the field could develop more effective methodological standards for replication are discussed.

replication crisis | reproducibility | p-hacking | researcher degrees of freedom | null hacking

In recent years, the field of psychology has suffered a crisis of confidence as investigators have reported failures to replicate many of its important original findings (1–7). This crisis has affected public perceptions of the soundness of scientific results (8–10) and spurred major initiatives at top scientific journals to improve methodological and reporting standards and encourage more investigators to conduct replication tests (11–14).

A major focus of this increased attention has been the threat to scientific integrity posed by "researcher degrees of freedom"—questionable research practices on the part of original investigators (15). It is now widely appreciated that original investigators face a conflict between the desire for accuracy and the career incentive to discover statistically significant results. There seems to be a widespread implicit presumption, however, that investigators who undertake replication tests are not subject to similar conflicts, but there are good reasons to believe that they are (16, 17). As replicability and research integrity have become topics of increasing interest, failures to replicate important original findings have begun to be published in top journals (1, 4–6, 18), while successful direct replications of existing findings receive much less attention. As a result, replicating investigators face a conflict between accuracy and statistical (non-)significance that is analogous to the one faced by original investigators: publication, attention from colleagues, and impact on the field largely depend on finding results that conflict with the original paper (17).*

The one-sided focus on "p-hacking," the motivated pursuit of statistically significant results by original investigators, ignores (and arguably, contributes to) a new threat to research integrity posed by "null hacking," the motivated pursuit of null results by replicating investigators (16). The purpose of this article is to demonstrate that replicator degrees of freedom, defined as discretion exercised at 2 stages of the replication process—experimental design and data analysis—can cause replicating investigators to arrive at incorrect conclusions about the replicability of an original finding. As a result, such tests can seem to move a field toward greater clarity and truth when, in reality, they are doing the opposite. If the field's well-intentioned initiatives to improve the replicability of its original findings are not guided by careful scrutiny of replicators' methods, they could have an ironic and counterproductive effect: trading one sort of misleading research finding (false-positive original findings) for another (false-negative replication results). This is a bad trade because the latter sort of misleading finding undoes the field's hard-won progress toward improved scientific understanding.

Others have already made versions of the 2 general methodological points that we make here: that empirical conclusions often hinge on analytic choices that competent investigators can disagree about and that replication tests that deviate from the design of the original study in material ways can create the misleading impression that the original finding was a false positive (19–25). Here, we provide an analysis of one prominent ongoing replication debate that demonstrates, concretely and directly, the implications of these 2 methodological principles for the field's interpretation of the many ostensible failures to replicate that are already in the literature and for how replication tests should be conducted going forward.

The failure of many replication tests to adequately recreate important design elements of the original studies in question is perhaps the most widely discussed point of disagreement about the field's new emphasis on replication (11, 19, 21, 26–30).† To

Significance

We show that commonly exercised flexibility at the experimental design and data analysis stages of replication testing makes it easy to publish false-negative replication results while maintaining the impression of methodological rigor. These findings have important implications for how the many ostensible nonreplications already in the literature should be interpreted and for how future replication tests should be conducted.

Author contributions: C.J.B., D.S.Y., and J.M.O. designed research; C.J.B., D.S.Y., and J.M.O. analyzed data; and C.J.B. and D.S.Y. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: All data and analysis scripts are posted on the Open Science Framework at https://osf.io/y5wsb/. An R package for conducting specification-curve analysis was developed for this work and is available for download at https://github.com/jmobrien/SpecCurve.

[1]To whom correspondence may be addressed. Email: christopher.bryan@chicagobooth.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910951116/-/DCSupplemental.

*Some might argue that the results-blind "Registered Replication Report" format offered by some journals (12) is an exception to this principle because editorial decisions are made before data are collected. Journals then publish the results regardless of the outcome. However, the career value of publishing successful replications (in terms of notoriety, hiring, and promotion at top institutions) still seems unlikely to be nearly as great as the career value of failed replications because the latter overturn settled knowledge and tend to attract more attention.

†Recent large-scale "Many Labs" replication projects have reported that little variation in replication effect size estimates is attributable to variation in certain arbitrarily chosen contextual factors (e.g., participating in person vs. online, Western vs. non-Western cultural settings) (30, 31). This has been interpreted as evidence that deviations from the original design are not an important cause of apparent failed replications (29–31). This interpretation of the Many Labs results reflects a misunderstanding of the context sensitivity point. Few would suggest that any deviation from the original context is likely to be important; rather, the point is that many effects are sensitive to particular contextual factors that the relevant theory suggests are important.

those who believe that such design factors are an important cause of what appear to be failed replications, the deviations in many replication tests can seem glaring. For example, a study of charitable giving that was the subject of a prominent failed replication involved mailing letters requesting donations to a Korean charity to compare the effect of different message types. Averaging across experimental groups, 11.8% of participants in the original study made donations in response to the letters, suggesting enough engagement with the letters to provide a real test of the effect of the message manipulation (32). A published failure to replicate that study's findings (4) involved sending a mass email and received donations, averaging across groups, from fewer than 0.002% of recipients, suggesting that the replication experiment failed to recreate the basic level of engagement with any of the messages that would be necessary to obtain a meaningful test of the message manipulation.

To those who are skeptical that deviations in design are an important explanation for failed replications, however, such arguments can seem like post hoc excuses from people reluctant to accept a difficult truth (26, 28–30, 33, 34). In dismissing these context-based arguments, some have issued a challenge to authors who claim that contextual variation can cause apparent failures to replicate: original investigators should conduct a new study correcting the design problems that they see in a failed replication test and show that this allows the original effect to replicate (28, 29, 33). Here, we provide stronger evidence than that. We show that, when the replicating investigators (i.e., independent researchers who are on record as being skeptical of the original effect) conduct a second replication test (35) that corrects a material deviation from the design of the original studies (18, 23, 33), they successfully replicate the original finding. The replicating investigators' choice, in this case, to address the original authors' context-based critique in a second replication study affords a rare opportunity to assess whether correcting such a contextual deviation can result in a successful replication. This question is directly relevant to a large and growing number of replication debates in the field (19, 22, 24, 25, 36–41).

A second class of replicator degree of freedom—flexibility about how to analyze data from replication tests (as opposed to original hypothesis tests)—has received much less attention in the academic discussion about research integrity. This is a serious oversight. Published reports of replication tests typically rely on a primary model specification, often complemented by "robustness checks"—a small number of variations on the primary specification that are meant to establish that the result of the main model is not dependent on arbitrary analytical choices (2, 5, 7, 42, 43). The problem is that replicating authors are free to choose both the primary model specification and which variations on that primary specification to report as robustness checks. In contrast to p-hacking by original investigators, even preregistration may not provide an adequate check against this problem. This is because it is possible to preregister numerous decisions that create the impression of scientific rigor but can be counted on to reduce the likelihood of statistically significant results. These include, for example, adopting an unnecessarily conservative significance criterion or method of estimating SEs, failing to adequately correct for chance failures of random assignment, or specifying models with such a large number of covariates or with covariates that are so highly correlated with the variable of interest that estimates become highly unstable (refs. 44, p. 390, 45, and 46). Although it is well known that some analytic choices can disproportionately favor the null hypothesis, debates about correct model specifications rarely end with consensus because competent researchers can reasonably disagree about analytic decisions (47), and parties on either side of a disagreement tend to select models that support their preferred result (48). Recently developed statistical methods (48, 49), however, provide effective ways to adjudicate such disagreements, making it possible to demonstrate the problem of replicator degrees of freedom in data analysis more definitively than was possible without these methods.

Why is a direct demonstration of these issues necessary? There are many examples in psychology and other behavioral sciences of conceptual arguments pointing out serious problems with a field's current methods or understanding that failed to produce meaningful change (19, 21, 50–53). In cases where fields have been persuaded that a problem is important enough to change their practices, it has typically been in response to a vivid and direct demonstration of how those practices can lead to incorrect conclusions (15, 54–56). That is, fields typically need to be shown just how badly an existing practice could lead them astray before they are persuaded enough to overcome the inertia maintaining the status quo. This analysis is important, in part, because it provides the sort of direct demonstration that has the potential to spur change.

In raising these issues, we do not intend to characterize the motives or practices of all or most investigators conducting tests of the replicability of original findings. Rather, we point out that there is a clear incentive for replicators to engage in such problematic behavior and that the field is vulnerable to publishing such misleading replication tests by showing that it already has.

Because we are not impartial arbiters of the debate in question,[‡] we take extraordinary precautions, in reanalyzing these replication data, to prevent our perspective on the substantive debate from influencing our results. We use 2 separate analytical approaches, each designed with the express purpose of maximizing transparency and minimizing the influence of arbitrary analytical choices (48, 49). Thus, an additional contribution of this paper is to show how 2 readily available analysis methods can be used to adjudicate disagreements about model specifications in replication tests.

Generalizability of This Debate

How generalizable is this analysis of a single replication debate to other such debates? This question is analogous to the question of whether an experimental procedure has operationalized the broader theoretical constructs at issue in ways that are likely to generalize beyond the specific experimental context. That is, just as investigators who use a laboratory procedure to study broad questions of human psychology must justify that their procedure adequately represents the real-life situation(s) of interest, we must justify that the key features of this debate are typical of the larger universe of published replication tests.

Three key features of this debate are typical of replication debates in psychology in recent years: 1) the replication tests were conducted in much larger samples than the original experiments, 2) the replication tests used experimental materials that were substantively identical to those used in the original experiments, and 3) the replicating authors reported several robustness checks on their main analysis. Replicating authors often rely heavily on these 3 features to support the validity of failures to replicate. Larger samples are often emphasized to suggest that failures to replicate should be given more weight than the original studies that they failed to replicate (1, 5, 6). The other 2 features are emphasized by replicators to address the very methodological concerns about replication that we raise in this paper. That is, the use of identical experimental materials is a common way to address questions about whether a replication test has successfully recreated the design of the original study (1, 5, 6). Additionally, robustness checks are often reported by replicating authors to allay any concern that a failure to replicate

[‡]We do not view this as a negative thing—few if any impartial investigators would have gone to the trouble to scrutinize these replication tests closely enough to identify their problems.

might be due to the particular analytical choices that those authors made (2, 5, 7, 42, 43). This reanalysis provides a critical test of whether these features of replication studies are, in fact, sufficient to prevent replicator degrees of freedom from influencing replication results.

The Debate in Question

We focus here on the debate about whether a subtle manipulation of language—referring to voting in an upcoming election with a predicate noun (e.g., "to be a voter") vs. a verb (e.g., "to vote")—can increase voter turnout. We provide a brief overview of the debate so far for context and to illustrate how the features of this debate are broadly representative of replication debates in the field.

Original Finding. Bryan et al. (57) published the findings of 2 field experiments in which participants who were randomly assigned to complete a brief online preelection survey on their thoughts and attitudes about "being a voter" in an upcoming election subsequently voted at a significantly higher rate (as measured using official voter turnout records) than participants who instead were assigned to answer questions about "voting." The authors argued that this is because noun wording frames the prospect of voting as an opportunity to claim a valued identity—to see oneself as the kind of person who votes (58, 59). In the period leading up to a high-profile election, when speculation about what "voters" will decide on Election Day is a major focus of both news media and casual conversation, the identity "voter" is expected to be highly valued, and the prospect of claiming it is expected to be particularly motivating.

Initial Replication Test. In 2016, Gerber et al. (18) published an article titled "A field experiment shows that subtle linguistic cues might not affect voter behavior," which described an experiment that bore superficial similarity to the original 2011 studies in that it used the same manipulation questions as the original study in a preelection survey. This replication study had a much larger sample than either of the original studies (more than 14 times the combined sample of the original 2 experiments) and found no evidence of any effect on turnout. The authors of that replication test acknowledged that their study design differed from that described in the original paper but dismissed the possibility that the differences were substantial enough to explain the different results (18, 33). Instead, they characterized their results as raising serious doubt about the replicability—or at least the robustness—of the original finding, suggesting that the 2011 finding "may have been a false positive. . ." (ref. 18, p. 7113).

Deviations in Replication Test from Original Experimental Context. Bryan et al. (23) pointed out in response that the replication test was conducted in a context in which the theory would not have predicted that the effect should occur. The original experiments were conducted the day before and morning of 2 high-profile elections: the 2008 US presidential election and the 2009 New Jersey gubernatorial election, both of which were major events that received substantial popular attention and coverage in the national media. By contrast, the replication study was conducted in the 4 d leading up to August primary elections for the 2014 midterms in 3 states (Table 1). Nearly half of those primaries were uncontested and therefore were not actually elections in any meaningful sense. Fewer than 10% of the primaries included in the replication test were competitive enough that they could plausibly have gotten significant attention from the public (23). Bryan et al. (23) demonstrated the psychological significance of this difference in context by asking participants to imagine either a tightly contested gubernatorial election (like the 2009 election in New Jersey) or an uncompetitive congressional primary. Participants indicated that "being a voter" would have far more important and positive identity implications in the former context than in the latter (23). The replicating authors dismissed these criticisms in a reply titled "Reply to Bryan et al.: Variation in context unlikely explanation of nonrobustness of noun versus verb result" (33).

The replication test attracted attention even outside of academia. It was featured, for example, in an article in *The Atlantic* by award-winning science journalist Ed Yong titled "Psychology's 'simple little tricks' are falling apart" (10). In that article, Yong (10) poses the question: "If it is so hard for teams of experienced and competent social scientists to get these techniques to work, what hope is there for them to be used more broadly?"

Second Replication Test in Appropriate Context. Roughly 2 y after the debate over the first replication test, Gerber et al. (35) published a second replication test conducted in the context of 4 relatively high-profile elections, correcting the first replication test's most serious deviation from the original experimental context. The most obvious of a number of remaining deviations from the original design in this second replication test (Table 1) was that it was conducted over 4 d—beginning 3 d before Election Day and ending when polls closed on Election Day. In the original experiments, the manipulation was administered only the day before and early (before 9 AM) the morning of Election Day. Therefore, the choice to administer the manipulation to some participants several days before the election represents a substantial departure from the original experiments, and data collected on those days should not be construed as a replication but rather as an extension of the original work. Moreover, the decision of Bryan et al. (23) to end data collection in the original experiments early the morning of Election Day (rather than continuing until the close of polls as the replicating authors did) was based on the logic that a boost in motivation to vote can only translate into actual voting if one still has time to get to the polls. As Election Day progresses, it becomes less and less likely that people will be able to add an unplanned and potentially time-consuming errand to their schedules. The effects of the replicating authors' choice to collect data until the close of polls were likely compounded by their failure to screen out prospective participants who had already voted. In sum, the Election Day sample in this second replication test likely included a large number of people who were treated after they had already voted and a large number of additional people who participated so late in the day that they were unlikely to be able to get to their polling place before it closed.

The replicating authors reported numerous and varied model specifications in their published report. Despite a seemingly thorough analysis and the substantial design improvements in this second replication test, they report having found no evidence that noun wording increased turnout relative to verb wording.

The Present Analysis

This paper presents a reanalysis of the data from the second replication test (35).[§] It is prompted by our observation that nowhere in that 2018 paper do the replicating authors report a test of the most straightforward, direct replication of the original experiments: was there an effect of noun wording on turnout among people who participated either the day before or early the morning of Election Day? The replicating authors do report models that include only data from the day before combined with the entire day of the election (until polls closed), which they characterize as a direct replication of the original experiments, but this is incorrect for the reasons noted above. Since the data from this replication test do not include the time of day at which the manipulation was administered, the closest approximation of

[§]Those data are available for download at https://huber.research.yale.edu/writings.html.

Table 1. List of the potentially important design features in the original experiments by Bryan et al. (57) that the replicating authors chose to deviate from in one or both of their published replication tests (18, 35)

No.	Design features	Original experiments (57)	Replication test #1 (18)	Replication test #2 (35)
1	Medium used to administer noun vs. verb manipulation	Online survey	**Telephone survey**	Online survey
2	Screened out prospective participants who had already voted?	Yes	**No**	**No**
3	Screened out nonnative English speakers?	Yes	**No**	**No**
4	Participants treated on Election Day until close of polls?	No	No	**Yes**
5	Participants treated on Election Day only before 9 AM?	Yes	**No**	**No**
6	Participants treated 1 d before Election Day?	Yes	Yes	Yes
7	Participants treated 2 d before Election Day?	No	**Yes**	**Yes**
8	Participants treated 3 d before Election Day?	No	**Yes**	**Yes**
9	Participants treated 4 d before Election Day?	No	**Yes**	No
10	Salient election context?	Yes: US presidential general and gubernatorial general (NJ)	**No: mostly uncompetitive midterm primaries (MI, MO, and TN)**	Yes: gubernatorial general (LA, MS, and KY) and mayoral general (Houston)

Design deviations are shown in bold font.

the original studies includes only data from the day before the election.

A preliminary analysis of the data from just the day before the election revealed that many of the most obvious model specifications yielded significant replications of the original noun-vs.-verb effect. We use one-tailed hypothesis tests because this experiment was a test of the replicability of a published finding, so there is an unambiguous directional prediction on record. Moreover, a leading methodology text for field experiments by the first author of both of these replication tests (60) recommends using one-tailed tests for "therapeutic interventions," which are designed to produce a positive effect, so the null hypothesis is that they have either no effect or a negative one (ref. 60, p. 64). In that book, the authors specifically apply this principle to one of their own voter turnout intervention experiments that, like the noun-wording intervention, was designed to increase but not decrease turnout (ref. 60, p. 158).

Using a one-tailed test and a simple ordinary least squares regression of turnout on experimental condition the day before the election (without covariates), we found a significant positive result ($b = 0.058$, $P = 0.036$, one tailed) that replicates the original finding. Adding a covariate indicating which of 2 survey firms the data came from, which many argue is necessary because randomization was implemented separately by the 2 firms (60, 61), the result is substantively the same ($b = 0.059$, $P = 0.032$, one tailed). Adding covariates for gender and race (white vs. nonwhite), because these tend to be correlated with voter turnout and can therefore improve the precision of the estimate, the P value drops further ($b = 0.067$, $P = 0.017$, one tailed).

A closer examination of the analyses reported by Gerber et al. (35) in support of their claim of nonreplication revealed that the replicating authors chose to include 3 features in all of their reported models that in combination are known to increase the risk of misleading results. They are the inclusion of a large number of covariates, substantial collinearity among covariates, and the inclusion of interaction terms. The combination of these 3 features can undermine the reliability of coefficient estimates by making them highly susceptible to influence from a small number of observations—an issue known as the "multivariate outlier problem" (44, 62).

Strictly speaking, covariates are not needed at all in a randomized experiment except to account for any stratification of random assignment built into the design (i.e., groups in which randomization was implemented separately, as with the 2 private survey firms that the replicating authors employed for data collection). Including covariates, however, can serve 2 additional legitimate purposes in randomized experiments: 1) they can be used to reduce any potential bias in treatment effect estimates that might result from chance failures of random assignment, and 2) they can be used to improve the precision of treatment effect estimates (i.e., reduce SEs) by explaining what would otherwise be treated as random variation in the outcome variable. These principled reasons to include covariates must be weighed against the risks of including a large number of covariates that have the problematic characteristics described above.

In their published models from the second replication test, Gerber et al. (35) included covariates for survey firm (which is necessary to account for stratified random assignment), the state in which data were collected, the date on which data were collected, the number of days before Election Day on which data were collected, gender, race, whether participants had voted in 15 previous elections, and the interaction between the 3 indicator variables for state and each of the other covariates in the model. Many of the voting history covariates were highly collinear (e.g., 17 correlations at 0.5 or above). Moreover, the inclusion of interaction terms between those highly correlated covariates and the state indicators resulted in a large number of additional covariates that were highly collinear. In all, the authors' model specification included 120 covariates. Twenty-five of those were so extremely collinear (or multicollinear) with other variables in the specified model that the software automatically excluded them because it was unable to estimate coefficients for them. The resulting model included 95 covariates and a very high degree of multicollinearity (*SI Appendix* has details).

The DFBETA statistic (44, 62, 63) provides the most direct indication of whether these problematic features of the replicating

authors' models, in fact, undermined the reliability of their results. DFBETA measures how sensitive a given coefficient estimate (in this case, the treatment effect) is to influence from a small number of observations (i.e., participants) in the data—a known risk of including a large number of highly correlated covariates (44). A widely recommended standard is that a DFBETA above $2/\sqrt{N}$ indicates that an observation is having a problematically large influence on the relevant coefficient estimate (44, 62, 63). We obtained DFBETA statistics for the treatment effect in the replicating authors' main model (using their publicly posted analysis script). Roughly 7% of observations (232 of the 3,078 participants) had DFBETA values above the recommended cutoff, suggesting that they were exerting a problematically high degree of influence on the treatment effect estimate in that model. This does not indicate that those observations are inherently problematic. In the simple models described above, controlling only for survey firm or only for survey firm, race, and gender, no cases have DFBETA levels above the recommended cutoff. Rather, the replicating authors' inclusion of so many multicollinear covariates seems to have caused their treatment effect estimate to be overly dependent on a small subset of observations.

On this basis, one can conclude that the model specifications reported by the replicating authors in their published report (35) are flawed and that conclusions based on them are unreliable. We noted above, however, that study results often hinge on data analytic decisions about which reasonable and competent researchers can disagree (15, 47, 48, 64, 65). Therefore, we used an analytical approach that is expressly designed to provide a comprehensive assessment of whether study data support an empirical conclusion when the influence of arbitrary researcher decisions on results is minimized. Our goal was to determine whether the statistical tests reported by Gerber et al. (35) accurately reveal the findings of their replication experiment or whether their study in fact replicated the original noun-vs.-verb effect on turnout but that finding was obscured by the authors' choice of particular model specifications.

If the latter of these possibilities were found to be the case, this would provide an important demonstration of how replicator degrees of freedom both at the data analysis stage and at the experimental design stage can result in the publication of misleading replication reports. That is, if, when analyzed appropriately, the second replication test by Gerber et al. (35) found evidence for the original effect, this would also show that an independent team of researchers was able to replicate the effect after they addressed the most serious deviations from the context of the original experiments that were present in their first replication test (33).

Treatment Effect Heterogeneity Due to Replication Design Choices. In addition to examining the level of support, in the data, for an overall effect of noun wording on turnout, we will assess whether one of the design choices exercised by Gerber et al. (35) in their second replication test affected the study's results. In the original experiments by Bryan et al. (57), the noun-vs.-verb manipulation was administered the day before or early the morning of Election Day. The replication experiment, by contrast, was conducted over a 4-d period beginning 3 d before Election Day and ending when polls closed (Table 1). After establishing whether the replication study found an overall effect of noun wording on turnout, we will examine more closely whether and how the treatment effect might have differed as a function of which day participants were treated on.

Specification-Curve Analysis. The primary statistical approach that we use, called "specification-curve analysis," involves running all reasonable model specifications (i.e., ones that are consistent with the relevant hypothesis, expected to be statistically valid, and are not redundant with other specifications in the set) (48). We then report the results in a format that makes transparent the effect of various analytical choices on the results. An associated significance test for the specification curve, called a "permutation test," quantifies how strongly the specification curve as a whole (i.e., all reasonable model specifications taken together) supports rejecting the null hypothesis. A permutation test involves generating a large number of simulated datasets that are exact copies of the real data except that the condition variable is randomly shuffled so that it can only be associated with the outcome variable by chance. Then, the specification curve is rerun in each of those simulated (null) datasets. The P value for the permutation test is simply the percentage of simulated, null datasets in which the specification curve yields results at least as extreme as the real data (e.g., with as many statistically significant results or with as large a median effect size estimate). That is, running the specification curve in a large number (e.g., 10,000) of simulated datasets in which the systematic "effect" of condition is known to be null provides a direct and precise estimate of the probability that the specification curve could have yielded as many statistically significant results (or a median effect size estimate as large) by chance as it did in the real data (48).

This method requires that researchers make judgments about what constitutes an exhaustive list of reasonable model specifications, but those decisions are articulated explicitly and the extent to which a result hinges on any given decision is easily determined from the resulting specification curve. We conduct this analysis in 2 stages. First (stage 1), we include only specifications about which we believe unbiased, methodologically informed researchers could reasonably agree or disagree (e.g., which covariates to include, which subsets of days to include). Second (stage 2), we expand the set of specifications to include ones that we believe are less defensible but that represent decisions that the replicating authors made in their paper (e.g., including a large number of highly collinear covariates, including data collected in the later part of Election Day, when many participants likely had already voted before participating or participated with insufficient time left to go vote before polls closed) (*SI Appendix*, Table S1 has a complete list of these decisions and the rationale behind them).[¶]

Results

The stage-1 specification curve contains 270 different regression models across 3 d before the election. We use 2 metrics to quantify the level of support for rejecting the null hypothesis across the full set of specifications: statistical significance (percentage of specifications for which $P < 0.05$, one tailed) and median estimated effect size. The statistical significance criterion asks, of the 270 different models included in this specification curve, how many individual models yield statistically significant support for rejecting the null hypothesis. Even if the null hypothesis were correct, that number would not likely be zero because some models would be expected to yield statistically significant results by chance. So, the permutation test provides a direct estimate of whether the observed percentage of statistically significant models, of the total of 270 models in the specification curve, is greater than one could plausibly observe if the null hypothesis were true. In this case, 137 (50.7%) of the 270 models yielded a P value below 0.05 (one tailed). The permutation test revealed that this is a much higher rate of statistical significance than would be expected if the null hypothesis were true. Using 10,000 simulated samples in which the null hypothesis is known to be true, the specification curve produced at least 137 statistically significant results only 3.1% of the time. This

[¶]An R package for conducting the specification curve analysis is available at https://github.com/jmobrien/SpecCurve (66).

percentage gives us the P value for the specification curve as a whole: if the null hypothesis were true in the real data, the specification curve would yield at least 137 statistically significant models only 3.1% of the time ($P_{\text{specification curve}} = 0.031$ by the statistical significance metric).

The second key metric that we use to quantify the strength of the specification curve results is median effect size. If the null hypothesis were true, the real effect size would be so close to zero that we would not be interested in it. Again, the permutation test provides a direct estimate of whether the median effect size estimate, across all models in the specification curve, is larger than one could plausibly observe if the null hypothesis were true in the real data. In this case, the median effect size estimate across the 270 models in the specification curve was that noun wording increased voter turnout by 3.6 percentage points. Practically, this is a large and meaningful estimated effect—especially considering how inexpensive the noun-wording treatment would be to administer at scale (*Discussion: This Replication Debate* has details). Of the 10,000 simulated samples generated for the permutation test, only 1.5% yielded a median effect size estimate as large or larger than the real data. This percentage again can be interpreted as a conventional P value that applies to the specification curve as a whole: if the null hypothesis were true, the specification curve would yield a median effect size this large only 1.5% of the time ($P_{\text{specification curve}} = 0.015$ by the effect size metric). In sum, whether we use the statistical significance or the effect size metric, the results provide clear support for the effect of noun wording on turnout.

The stage-2 specification curve expands the set of specifications to include 930 additional models that represent decisions that Gerber et al. (35) made in their published paper (*SI Appendix*, Table S1). The expanded specification curve included a total of 1,200 different models, including ones that use data from participants treated on Election Day, when many participants are expected either to have already voted or to have had very little time left before the close of polls after completing the manipulation. Of those 1,200 models, 507 (42.2%) yielded a result with a one-tailed P value less than 0.05. In a permutation test, using 10,000 simulated null samples, only 3.4% yielded as many or more significant results ($P_{\text{specification curve}} = 0.034$ by the statistical significance metric).

The median effect size estimate across all 1,200 models in the stage-2 specification curve is that noun wording increased turnout by 2.9 percentage points, still a large estimated effect (e.g., this is slightly larger than the effect of the costly method of having live volunteers speak on the phone with each prospective voter) (67). The permutation test reveals that, of 10,000 simulated samples in which the null hypothesis is true, only 2.4% yielded a median effect size estimate as large or larger than that ($P_{\text{specification curve}} = 0.024$ by the effect size metric). In sum, the evidence across the expanded set of model specifications that includes the main analytical choices by Gerber et al. (35) supports a substantial and robust effect consistent with the original finding by Bryan et al. (57). This again was true whether we used the percentage of statistically significant results or the median effect size across all specifications as the criterion in the permutation test. Perhaps the most striking finding from the specification-curve analysis is that, in every subset of the data that we examined, the model specifications using the exact set of highly collinear covariates used by the replicating authors (35) were dramatic outliers, yielding point estimates that were much lower than any other model specification (Fig. 1).

Treatment Effect Heterogeneity Due to Replication Design Choices

The results described so far make clear that noun wording had a significant effect on turnout overall in the experiment by Gerber et al. (35). However, the specification curve results also strongly suggest that the replicating authors' data analysis choices might not be the only replicator degrees of freedom influencing results. Rather, the replicating authors' design choices regarding the window of time in which the study was conducted may have further driven the treatment effect estimate downward (Table 1). Unsurprisingly, the effect seems weak or nonexistent among participants who completed the manipulation on Election Day (Fig. 1 *A*, *B*, and *C* vs. *D*, *E*, and *F*, respectively). As we have noted, many who participated that day would already have voted before they were treated, and many others would have been treated so close to the close of polls that they would not have had time to go vote afterward. More interestingly, while the effect seems strong and robust among participants who were treated 1 or 2 d before Election Day, it seems to drop off sharply among participants who were treated 3 d before the election (Fig. 1 *B* vs. *C* and *E* vs. *F*). This could indicate a boundary on how long the motivational effects of noun wording can last. If this apparent heterogeneity were reliable, it 1) would suggest that including data from Election Day (which in this study, means all day until the polls closed) and from 3 d before the election is artificially depressing the treatment effect estimate and 2) would provide another demonstration of how replicator degrees of freedom at the design stage can produce misleading replication results.

Gerber and Green (ref. 60, p. 310) recommend using machine-learning approaches to automate the search for systematic sources of heterogeneity in treatment effects that are not predicted in advance by theory. Because such approaches are automated, minimizing the influence of researcher decisions on results, they are not subject to the concerns about multiple hypothesis testing that undermine the credibility of most post hoc discoveries of moderation (68).

Bayesian Causal Forest. In line with the suggestion of Gerber and Green (ref. 60, p. 310), we used a Bayesian machine-learning algorithm, called Bayesian Causal Forest (BCF), that has been shown repeatedly by both its creators and other leading statisticians to be the most effective of the state-of-the-art methods for identifying true, systematic sources of treatment effect heterogeneity while avoiding false positives (49, 69, 70). That is, in 2 open head-to-head competitions, BCF was found to be superior to other cutting-edge methods at differentiating true, systematic heterogeneity in treatment effects from random variation (49, 69, 70).

Because of the replicating authors' design choices, we exclude the data collected on Election Day, when an unknown but likely substantial proportion of participants was treated either after their vote had already been recorded or with so little time left in which to vote that the manipulation could not plausibly have influenced whether they voted. We do this because BCF imposes a penalty on the estimated probability that systematic heterogeneity exists for every combination of study days that it considers. In this case, the number of possible combinations jumps from 7 to 15 when considering 4 study days instead of 3. Therefore, including data from Election Day would cause BCF to impose a penalty for considering 8 possible combinations of days that we know not to be valid tests of the hypothesis.

The BCF analysis confirmed that the effect of noun wording on turnout is systematically stronger among participants who were treated either 1 or 2 d before the election than it is among participants who were treated 3 d before the election (posterior probability of systematic heterogeneity 0.96 or 24:1 odds).[#] The BCF results also confirm that there is a systematic positive effect

[#]BCF is a Bayesian algorithm so its results are described in terms of posterior probabilities, which should not be confused with P values. This Bayesian posterior means that, even beginning with a strong presumption that there is no moderation in the treatment effect, the pattern in the data is so clear and strong that the algorithm conservatively estimates a probability of 0.96 that there is true, systematic heterogeneity in the treatment effect based on when participants were treated.

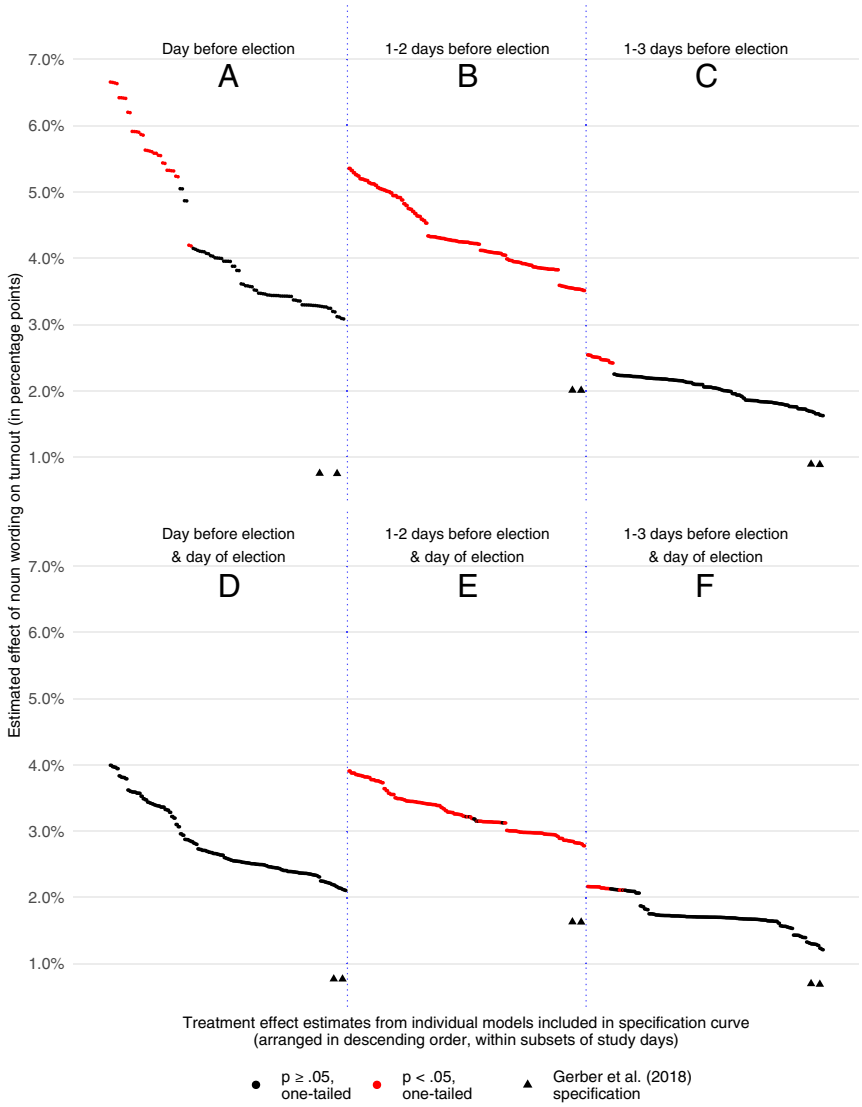

Fig. 1. Plot of the effect size estimates from 1,200 models in the stage-2 specification curve, which includes the 270 specifications from the stage-1 specification curve and 930 models that represent analytical choices made by the replicating authors. *A*–*F* represent subsets of study days. The triangular points along the bottom of each panel indicate specifications using the exact set of 95 covariates used by Gerber et al. (35).

of noun wording on turnout among participants who were treated either 1 or 2 d before the election (posterior probability of 0.95 or 19:1 odds). By contrast, BCF finds little or no evidence of a systematic effect of noun wording on turnout among participants who were treated 3 d before the election (posterior probability = 0.51); that is, it is roughly equally likely that there is no treatment effect on that day as it is that there is an effect (*SI Appendix*, Fig. S1 has details of BCF analysis).

These results confirm that the replicating authors' choice to begin administering the treatment to participants as early as 3 d before the election (35) was suppressing the treatment effect estimate. To obtain the specification curve estimate of the treatment effect removing this suppressing factor, we looked at its results using only data from 1 or 2 d before the election—the days on which the BCF analysis indicates that there was an effect.

Specification Curve Removing the Suppressing Effect of Replication Design Choices. We focus on the more conservative stage-2 specification curve, which includes the analytical decisions that Gerber et al. (35) made in their published paper. (The stage-1 curve results were substantively identical.) When limiting to specifications that include only data collected 1 or 2 d before Election Day, the stage-2 specification curve contains 218 different models, of which 216 (99.1%) yield a one-tailed *P* value less than 0.05. The only 2 specifications in this set that do not yield a statistically significant result are the 2 that use the exact set of highly collinear covariates reported by the replicating authors (35) (Fig. 1*B*). In a permutation test, only 1.0% of the 10,000 simulated null samples yielded as many or more statistically significant results ($P_{\text{specification curve}} = 0.010$ by the statistical significance metric).

The median effect size estimate in this specification curve was 4.2 percentage points. In a permutation test, only 0.8% of the 10,000 simulated null samples yielded a median effect size estimate as large or larger than the real data ($P_{\text{specification curve}} = 0.008$ by the effect size metric). In sum, virtually any model specification that one uses to estimate the effect of noun wording on turnout in the 2 d prior to Election Day yields statistically significant support for the hypothesis and a substantial effect size estimate.

Discussion: This Replication Debate

The clear conclusion of this reanalysis is that the replication test by Gerber et al. (35) yielded strong evidence in support of a noun-wording effect on turnout. That conclusion is consistent across 2 different statistical approaches to hypothesis testing, both of which are designed to provide a comprehensive assessment of whether an empirical conclusion is supported by the data while minimizing the influence of researcher decisions on results. It should be noted that this reanalysis was facilitated by the replicating authors' choice to make their data publicly available. Although the conclusion that those authors reached (35) is not supported by their data, their decision to provide open access to those data allowed us to identify and correct the error.

Noun wording had a positive effect on turnout specifically when it was administered 1 or 2 d before Election Day. On those days, the median estimated effect of noun wording on turnout in the specification-curve analysis was quite large: 4.2 percentage points. In evaluating the importance of an effect of this size, it is useful to consider how it compares with alternative methods of increasing turnout. Past research provides a number of useful benchmarks. The most effective conventional method of increasing turnout (but also the most costly by far) is door-to-door canvassing, which increases turnout among those who actually answer the door and talk to the canvasser by 6.7 percentage points, on average (67). The second most effective method is live telephone calls from volunteers urging people to vote, which are also quite costly and increase turnout by roughly 2.8 percentage points among those who answer the phone and talk to the caller (67).[‖] Considering these benchmarks and the low cost of the noun-wording treatment, the 4.2-percentage point effect is remarkably large.

Even the large estimated effect of noun wording reported here, moreover, is likely conservative. Two replicator degrees of freedom exercised in the design phase of this experiment cannot be corrected in data analysis and likely suppressed the treatment effect for a substantial proportion of participants. First, in contrast with the original experiments by Bryan et al. (57), participants in this experiment were not asked whether they had already voted before being allowed to enroll in the experiment (Table 1, no. 2). This is important because 2 of the 4 elections that the replication test was conducted in had high rates of early voting. In the Houston mayoral election, 39% of all votes were cast before the study began, and an additional 11% of votes in that election were cast absentee and therefore were likely (but not necessarily) cast before the replication test began. In the Louisiana gubernatorial election, 23% of all votes were cast either early (before the start of the experiment) or absentee (the state does not keep separate records for those 2 categories of voting). This means that many participants in the replication test likely could not have been induced to vote by the treatment because they had already voted.

Second, because the noun-vs.-verb manipulation relies on people's ability to interpret the difference in meaning behind a subtle linguistic variation, the original experiments by Bryan et al. (57, 71, 72) only included native-English-speaking participants. The replication test did not assess participants' comfort in English (Table 1, no. 3), so some portion of participants in their study probably lacked the language proficiency to detect the subtle meaning behind the noun wording.

For these reasons, the median effect size estimate of 4.2 percentage points likely underestimates the effect of noun wording on people who have not already voted and who are able to detect the meaning behind noun wording. The second of the original experiments used a professionally managed panel of adults recruited via probability sampling methods and found an estimated effect of 10.9 percentage points (57). This is not to suggest that the replicating authors' failure to screen out early voters and nonnative English speakers suppressed what otherwise would have been an 11-percentage-point effect. The effect of noun wording on turnout almost certainly varies in size depending on the specific election context and the characteristics of the population in which it is implemented.

The results of this reanalysis do not mean, and we do not claim, that noun wording will increase turnout in all circumstances. Like most (if not all) psychologically informed interventions, this effect depends on a context that is amenable to the relevant psychology (73, 74). One critically important factor in producing that context is the nature of the election. The identity "voter" is only expected to be motivating to most citizens in a context where the decisions voters are about to make feel important and meaningful (23). Another important factor is the timing of the treatment (75). This analysis found that, in this experiment, the treatment was effective when administered 1 to 2 d before an election but not when it was administered 3 d before.** This is a valuable discovery that provides guidance to practitioners seeking to apply this intervention and provides the basis for new theorizing about the psychological and behavioral mechanisms by which a temporary boost in motivation to vote is likely to translate into real, consequential behavior later on (76). The fact that this discovery was missed by the replicating authors underscores the costs of the unnecessarily adversarial climate around replication in the field today and the need for a culture in which original and replicating authors both seek to understand the complexities of human behavior in context.

It is also important to distinguish between a psychological effect (e.g., the feeling of motivation) and its practical manifestation (e.g., the act of voting). Separate contextual factors are expected to influence whether each of these occurs. For example, even when the context is in place to allow noun wording to increase people's motivation to vote, other pragmatic factors (e.g., severe weather or a rise in gas prices that make it costlier or more difficult to get to the polls) might interfere with the translation of that boost in motivation into a systematic change in behavior.

General Discussion

Beginning roughly in 2011, in response to increased awareness of the potential for p-hacking to lead to high rates of false-positive results, there was a sudden and dramatic increase in the number of studies testing the replicability of psychology findings (3). This rush to scale up replication testing left little time for the field to have a thoughtful discussion or reach consensus about appropriate methodological standards for replication tests. Instead, replication investigators have tended to rely on informal rules of thumb to address critically important validity issues. As a result, a large number of replication tests have now been published, and there is widespread disagreement about how to interpret them or whether many of them are informative at all (19, 21, 28, 29, 34).

Using an in-depth analysis of data from one prominent and representative replication debate in the field, we demonstrate that 3 such rules of thumb—recruiting large samples, matching the experimental materials of the original study, and reporting robustness checks on primary statistical analyses—are insufficient for their purpose. The 2 replication tests in this debate contained serious validity problems on the very dimensions that

[‖]These are conservative benchmarks. Both the door-to-door canvassing and phone call effect size estimates refer to "complier average causal effects" (60), which are adjusted upward to account for the fact that many participants assigned to the treatment condition (i.e., to be contacted) are unreachable or refuse to talk with the canvasser or caller. All estimates of the noun-vs.-verb effect that we report in this reanalysis are intent-to-treat effects, which do not adjust for any failures to treat participants assigned to the treatment condition.

**Both of these factors are probably best thought of as necessary but not sufficient conditions. Understanding the necessary and sufficient conditions for this (or any) effect requires an accumulation of knowledge from a large number of studies in a range of contexts and populations using a variety of design specifications.

these rules of thumb are meant to address and were nevertheless published in well-respected scientific journals.

The publication of a highly implausible finding (i.e., the existence of precognition) in social psychology's flagship journal (77) served as an important impetus for the field's shift toward more rigorous methodological standards (3). Here, we demonstrate that 2 ostensible failures to replicate a prominent finding yielded misleading conclusions despite adhering to the methodological and reporting standards common to such tests. Our hope is that this demonstration prompts a shift toward more rigorous and appropriate methodological standards for replication tests. Such a change is needed before the field can hope to reap the full potential benefit of its increased emphasis on replication testing. This change is also urgent, as replicator degrees of freedom risk undermining the credibility of the broader research integrity movement, hobbling its effectiveness in guarding against real threats to scientific progress.

This demonstration also suggests that additional scrutiny of the many ostensible failed replications already in the literature is warranted. This analysis provides a model of how that scrutiny can be applied. Indeed, just in the time since this reanalysis was conducted, another published "failure to replicate" (42) has been shown to in fact have been a successful replication using the same combination of specification curve and BCF analyses (78).

In considering what the appropriate methodological standards for replication tests should be going forward, we suggest that the field should be mindful of 4 broad points supported by this analysis. First, independent replications should not be presumed to be unbiased. There is a clear incentive for replicators to obtain results that conflict with the original study (17) that should be presumed to exert as much influence on how replicating investigators exercise degrees of freedom as the incentive to find significant results influences original investigators. Even before this incentive can shape how investigators carry out replication tests, it might influence which effects they choose to replicate. That is, the incentive to find null results in replication tests could bias investigators to select effects for replication testing that they already believe are false.

Second, sample size alone is not a useful indicator of the accuracy of an effect size estimate. Both of the replication tests in this debate used samples that were many times larger than the original experiments, and both resulted in the publication of misleading results because the experimental design (18, 23) or the data analysis (35) was not appropriate. Additional statistical power does little to improve the accuracy of inferences from data if the other substantive aspects of the work are not executed correctly (79). Although this point is obvious, it is often missed. The authors of failed replication tests frequently argue (directly or implicitly) that studies with conflicting findings should be given weight in proportion to their sample size (1, 5, 6, 18). Leaders in psychology's replication movement have even suggested that academic hiring and promotion committees should treat the average sample size of studies published in a journal as a "superior way of quantifying journal quality" (80).

Third, those who have dismissed the argument that many ostensible failures to replicate might be attributable to a failure to recreate the context or psychological experience created in the original studies that they are replicating (10, 28, 29, 33, 34) should reconsider their position on this issue. Relatedly, reviewers and editors who evaluate ostensible failed replications should be skeptical of claims, direct or implied, that using the same experimental materials is sufficient to ensure a faithful replication test. Some have suggested that, if a psychological effect depends on seemingly subtle contextual factors, it is like a delicate plant, hopelessly fragile (10, 29). This argument reveals a failure of perspective taking. The insight that aspects of a context can seem trivial to some people but feel profoundly consequential to others is a founding principle of social psychology (81). With this in mind, we suggest an alternative to the misguided plant metaphor. A psychology experiment that depends on seemingly subtle contextual elements is more like a chain; a single poorly wrought link might cause it to fail, but this does not mean that chains are weak.

On this point, we must acknowledge that original investigators bear considerable responsibility for the misunderstanding. In published descriptions of our work, our field has tended to emphasize the surprising simplicity of our experiments and downplay the careful thought that goes into crafting many background elements of our experimental designs. (This may be especially true of social psychologists.) This practice makes our articles engaging to read but is incompatible with the scientific method; we must provide detailed information, in our published reports, about what our theories suggest are the likely boundary conditions on our findings (82). If authors find it awkward to integrate this information into their articles, it could be included in an appendix, for example, in the form of a "replication guide."

Fourth, robustness checks are not sufficient to protect against analysis degrees of freedom. Such checks give the impression that a result is robust to any reasonable variation on the primary model specification when, in reality, they only show that it is robust to the particular set of variations that the author chose to report. Even preregistration is an incomplete solution to the problem of null hacking in data analysis because, as we noted above, there are numerous analysis choices that one can make a priori that make null results more likely even if a systematic effect is present in the data. Fortunately, recent developments in data analysis make this a tractable problem. Perhaps the clearest, most concrete implication of this analysis is that specification-curve analysis should be standard practice in replication testing. For example, replicating authors could consult with original authors to develop a preanalysis plan, and any disagreements about reasonable choices for model specifications could be entered into a specification curve (or BCF). These methods will not necessarily resolve disagreements about model specification, but at a minimum, they will expose the analytical decisions on which a result hinges, facilitating a debate about the merits of different choices (48).

Finally, this demonstration underscores the need for the field to move away from the simplistic mentality that emphasizes whether an effect is "real" or not, or the myopic focus on estimating a "true average effect" without regard to how the details of a study's design or analysis can be important determinants of the size of the resulting treatment effect (75). The fact that the effects of psychological interventions are sensitive to context does not mean that they are hopelessly complicated to produce (10). Indeed, this demonstration shows that even investigators who are skeptical of an intervention effect are able to successfully replicate it by implementing even minimal design requirements for recreating the necessary context. In the words of prominent methodology expert Andrew Gelman, "Once we accept that treatment effects vary. . .[w]e move away from is-it-there-or-is-it-not-there to a more helpful, contextually informed perspective" (ref. 68, p. 636).

Materials and Methods

A summary of the key methodological details from the published replication report by Gerber et al. (35) is given here. Additional detail can be found in that published report.

Participants. Participants were recruited through 2 commercial operators of opt-in survey panels: Survey Sampling International (SSI) and YouGov. SSI panel members were admitted into the experiment if they reported that they were of voting age, were registered to vote at their current address, lived at a zip code that falls entirely or predominantly within 1 of the 4 target regions (*Procedure*), and provided a valid first and last name. YouGov panel members were admitted if they reported a zip code within the target regions and if their existing profile data with YouGov indicated that they met the same criteria. The sample comprises 3,078 participants.

Procedure. The experiment was conducted in the days leading up to the 2015 gubernatorial elections in Kentucky, Louisiana, and Mississippi and the mayoral election in Houston, Texas. All participants completed a 10-item online survey about their thoughts and attitudes about voting in the upcoming election. The content of the questions was identical in both conditions, but participants were randomly assigned to receive a version of the survey that referred to voting with either a verb or a predicate noun (e.g., "How important is it to you to [vote/be a voter] in Tuesday's election"). Participants were randomly assigned to be enrolled in the experiment 3 d before the election with 25% probability, 2 d before with 25% probability, and the day before or day of the election (until polls closed) with 50% probability.

Variables in Data File. In addition to indicators for experimental condition and the survey firm that collected the data, the public data file for this experiment contained indicators for the gender, race, and home state of each participant, date of participation, and whether participants had voted in each of 15 elections between 2004 and 2015 (inclusive). Finally, the data include what the authors described as "inverse probability weights," with no explanation, in the main article or the appendix, of how those were computed or what purpose they are meant to serve.

Data availability. All data and associated analysis code are available at https://osf.io/y5wsb/ (83).

ACKNOWLEDGMENTS. We thank C. Dweck, J. Bowers, J. Murray, N. Malhotra, J. Druckman, J. Krosnick, E. Dunn, J. Simmons, Z. Hajnal, P. Loewen, K. Pathakis, K. Matush, and L. Sanford for helpful comments or advice. Preparation of this manuscript was supported by a fellowship from the Center for Advanced Study in the Behavioral Sciences at Stanford University (to C.J.B.), a William T. Grant Scholars award (to D.S.Y.), the National Institutes of Health under award R01HD084772, and grant P2CHD042849, Population Research Center, awarded to the Population Research Center at The University of Texas at Austin by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

1. C. F. Camerer *et al*., Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
2. C. R. Harris, N. Coburn, D. Rohrer, H. Pashler, Two failures to replicate high-performance-goal priming effects. *PLoS One* **8**, e72467 (2013).
3. L. D. Nelson, J. Simmons, U. Simonsohn, Psychology's renaissance. *Annu. Rev. Psychol.* **69**, 511–534 (2018).
4. O. S. Collaboration; Open Science Collaboration, PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
5. E. Ranehill *et al*., Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychol. Sci.* **26**, 653–656 (2015).
6. D. Rohrer, H. Pashler, C. R. Harris, Do subtle reminders of money change people's political views? *J. Exp. Psychol. Gen.* **144**, e73–e85.
7. S. Doyen, O. Klein, C.-L. Pichon, A. Cleeremans, Behavioral priming: It's all in the mind, but whose mind? *PLoS One* **7**, e29081 (2012).
8. B. Carey, Many psychology findings not as strong as claimed, study says. *NY Times*, 27 August 2015. https://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html. Accessed 30 November 2018.
9. B. Carey, New critique sees flaws in landmark analysis of psychology studies. *NY Times*, 3 March 2016. https://www.nytimes.com/2016/03/04/science/psychology-replication-reproducibility-project.html. Accessed 30 November 2018.
10. E. Yong, A worrying trend for psychology's 'simple little tricks.' *Atlantic*, 8 September 2016. https://www.theatlantic.com/science/archive/2016/09/can-simple-tricks-mobilise-voters-and-help-students/499109/. Accessed 30 November 2018.
11. B. A. Nosek, D. Lakens, Registered reports: A method to increase the credibility of published results. *Soc. Psychol.* **45**, 137–141.
12. D. J. Simons, A. O. Holcombe, B. A. Spellman, An introduction to registered replication reports at perspectives on psychological science. *Perspect. Psychol. Sci.* **9**, 552–555 (2014).
13. M. R. Munafò *et al*., A manifesto for reproducible science. *Nature Hum. Behav.* **1**, 0021 (2017).
14. J. Berg, Progress on reproducibility. *Science* **359**, 9 (2018).
15. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
16. J. Protzko, Null-hacking, a lurking problem in the open science movement. https://psyarxiv.com/9y3mp/ (21 June 2018).
17. A. J. Berinsky, J. N. Druckman, T. Yamamoto, *Why Replications Do Not Fix the Reproducibility Crisis: A Model and Evidence Rom a Large-Scale Vignette Experiment* (Institute for Policy Research, Northwestern University, 2019).
18. A. S. Gerber, G. A. Huber, D. R. Biggers, D. J. Hendry, A field experiment shows that subtle linguistic cues might not affect voter behavior. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7112–7117 (2016).
19. D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, Comment on "Estimating the reproducibility of psychological science." *Science* **351**, 1037 (2016).
20. W. Stroebe, F. Strack, The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **9**, 59–71 (2014).
21. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6454–6459 (2016).
22. A. Norenzayan, Some reflections on the many Labs 2 replication of norenzayan, smith, kim, and nisbett's (2002) study 2: Cultural preferences for formal versus intuitive reasoning. *Adv. Methods Pract. Psychol. Sci.* **1**, 499–500 (2018).
23. C. J. Bryan, G. M. Walton, C. S. Dweck, Psychologically authentic versus inauthentic replication attempts. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6548 (2016).
24. E. Shafir, The workings of choosing and rejecting: Commentary on many Labs 2. *Adv. Methods Pract Psychol. Sci.* **1**, 495–496 (2018).
25. N. Schwarz, F. Strack, Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Soc. Psychol.* **45**, 305–306 (2014).
26. Y. Inbar, Association between contextual dependence and replicability in psychology may be spurious. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4933–E4934 (2016).
27. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4935–E4936 (2016).
28. S. Srivastava, "Moderator interpretations of the Reproducibility Project." *The Hardest Science*. https://thehardestscience.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/. Accessed 30 November 2018.
29. B. W. Roberts, "The new rules of research." *pigee*. https://pigee.wordpress.com/2015/09/17/the-new-rules-of-research/. Accessed 30 November 2018.
30. C. R. Ebersole *et al*., Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
31. R. A. Klein *et al*., Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
32. M. Koo, A. Fishbach, Dynamics of self-regulation: How (un)accomplished goal actions affect motivation. *J. Pers. Soc. Psychol.* **94**, 183–195 (2008).
33. A. S. Gerber, G. A. Huber, D. R. Biggers, D. J. Hendry, Reply to Bryan et al.: Variation in context unlikely explanation of nonrobustness of noun versus verb results. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6549–E6550 (2016).
34. C. J. Anderson *et al*., Response to Comment on "Estimating the reproducibility of psychological science." *Science* **351**, 1037 (2016).
35. 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. *Polit. Psychol.* **39**, 925–938 (2018).
36. M. J. Ferguson, T. J. Carter, R. R. Hassin, Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Soc. Psychol.* **45**, 301–302 (2014).
37. M. C. Frank, T. Holubar, Data from "Replication of Monin, Sawyer, & Marquez (2008, JPSP 95(1), Exp. 4)." Open Science Framework. https://osf.io/pz0my/. Accessed 13 May 2019.
38. S. W. S. Lee, N. Schwarz, Methodological deviation from the original experiment. *Nat. Hum. Behav.* **2**, 605 (2018).
39. D. C. Kidd, E. Castano, Panero et al. (2016): Failure to replicate methods caused the failure to replicate results. *J. Pers. Soc. Psychol.* **112**, e1–e4 (2017).
40. B. Sparrow, The importance of contextual relevance. *Nat. Hum. Behav.* **2**, 607 (2018).
41. M. A. Pyc, K. A. Rawson, The mediator effectiveness hypothesis revisited. *Nat. Hum. Behav.* **2**, 608 (2018).
42. Y. Li, T. C. Bates, You can't change your basic ability, but you work at things, and that's how we get hard things done: Testing the role of growth mindset on response to setbacks, educational attainment, and cognitive ability. *J. Exp. Psychol. Gen.* **148**, 1640–1655 (2019).
43. C. R. Dobronyi, P. Oreopoulos, U. Petronijevic, Goal setting, academic reminders, and college success: A large-scale field experiment. *J. Res. Educ. Eff.* **12**, 38–66 (2019).
44. J. Cohen, P. Cohen, S. G. West, L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates Publishers, ed. 3, 2003).
45. D. K. Ginther, S. Kahn, WOMEN IN SCIENCE. Comment on "Expectations of brilliance underlie gender distributions across academic disciplines." *Science* **349**, 391 (2015).
46. A. Cimpian, S.-J. Leslie, WOMEN IN SCIENCE. Response to Comment on "Expectations of brilliance underlie gender distributions across academic disciplines." *Science* **349**, 391 (2015).
47. R. Silberzahn *et al*., Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Adv. Methods Pract Psychol. Sci.* **1**, 337–356 (2018).
48. U. Simonsohn, J. P. Simmons, L. D. Nelson, Specification Curve: Descriptive and inferential statistics on all reasonable specifications. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694998. Accessed 18 July 2019.
49. P. R. Hahn, J. S. Murray, C. Carvalho, Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. arXiv:1706.09523 (29 June 2017).
50. D. F. Hendry, Econometrics-alchemy or science? *Economica* **47**, 387–406 (1980).
51. J. M. Keynes, Professor Tinbergen's method. *Econ. J. (Lond.)* **49**, 558–577 (1939).
52. N. L. Kerr, HARKing: Hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2**, 196–217 (1998).

53. D. O. Sears, College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *J. Pers. Soc. Psychol.* **51**, 515–530 (1986).
54. J. D. Angrist, J.-S. Pischke, The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *J. Econ. Perspect.* **24**, 3–30 (2010).
55. R. J. LaLonde, Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* **76**, 604–620 (1986).
56. E. Cohen-Cole, J. M. Fletcher, Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *BMJ* **337**, a2533 (2008).
57. C. J. Bryan, G. M. Walton, T. Rogers, C. S. Dweck, Motivating voter turnout by invoking the self. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12653–12656 (2011).
58. S. A. Gelman, G. D. Heyman, Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychol. Sci.* **10**, 489–493 (1999).
59. S. A. Gelman, M. Hollander, J. Star, G. D. Heyman, "The role of language in the construction of kinds" in *Psychology of Learning and Motivation*, D. L. Medin, Ed. (Academic Press, 2000), vol. 39, pp. 201–263.
60. A. S. Gerber, D. P. Green, *Field Experiments: Design, Analysis, and Interpretation* (W. W. Norton & Company, ed. 1, 2012).
61. R. Glennerster, K. Takavarasha, *Running Randomized Evaluations: A Practical Guide* (Princeton University Press, 2013).
62. D. A. Belsley, E. Kuh, R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (John Wiley & Sons, 1980).
63. K. A. Bollen, R. W. Jackman, Regression diagnostics: An expository treatment of outliers and influential cases. *Sociol. Methods Res.* **13**, 510–542 (1985).
64. E. E. Leamer, Let's take the con out of econometrics. *Am. Econ. Rev.* **73**, 31–43 (1983).
65. E. L. Glaeser, "Researcher incentives and empirical methods" (National Bureau of Economic Research, 2006), NBER Technical Working Paper no. 329.
66. J. M. O'Brien, R package for constructing specification curves. Github. https://github.com/jmobrien/SpecCurve. Deposited 23 September 2019.
67. D. P. Green, A. S. Gerber, *Get Out the Vote: How to Increase Voter Turnout* (Brookings Institution Press, ed. 3, 2015).
68. A. Gelman, The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *J. Manag.* **41**, 632–643 (2015).
69. 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 (9 July 2017).
70. T. Wendling *et al*., Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat. Med.* **37**, 3309–3324 (2018).
71. C. J. Bryan, G. S. Adams, B. Monin, When cheating would make you a cheater: Implicating the self prevents unethical behavior. *J. Exp. Psychol. Gen.* **142**, 1001–1005 (2013).
72. C. J. Bryan, A. Master, G. M. Walton, "Helping" versus "being a helper": Invoking the self to increase helping in young children. *Child Dev.* **85**, 1836–1842 (2014).
73. D. S. Yeager, G. M. Walton, Social-psychological interventions in education: They're not magic. *Rev. Educ. Res.* **81**, 267–301 (2011).
74. G. M. Walton, T. D. Wilson, Wise interventions: Psychological remedies for social and personal problems. *Psychol. Rev.* **125**, 617–655 (2018).
75. V. K. Alogna *et al*., Registered replication report: Schooler and engstler-schooler (1990). *Perspect. Psychol. Sci.* **9**, 556–578 (2014).
76. D. T. Miller, J. E. Dannals, J. J. Zlatev, Behavioral processes in long-lag intervention studies. *Perspect. Psychol. Sci.* **12**, 454–467 (2017).
77. C. S. Dweck, D. S. Yeager, Data from "Re-analysis commentary - 9-6-19.pdf (Version: 1)." Open Science Framework. https://osf.io/kvqh4/. Accessed 1 November 2019.
78. C. S. Dweck, D. S. Yeager, A simple re-analysis overturns a "failure to replicate" and highlights an opportunity to improve scientific practice. in press.
79. B. B. McShane, U. Böckenholt, You cannot step into the same river twice: When power analyses are optimistic. *Perspect. Psychol. Sci.* **9**, 612–625 (2014).
80. R. C. Fraley, S. Vazire, The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One* **9**, e109019 (2014).
81. L. Ross, R. E. Nisbett, *The Person and the Situation: Perspectives of Social Psychology* (McGraw-Hill, 1991).
82. D. J. Simons, Y. Shoda, D. S. Lindsay, Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).
83. C. J. Bryan, D. S. Yeager, J. M. O'Brien, Replicator degrees of freedom allow publication of misleading "failures to replicate." Open Science Framework. https://osf.io/y5wsb/. Deposited 1 November 2019.

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.

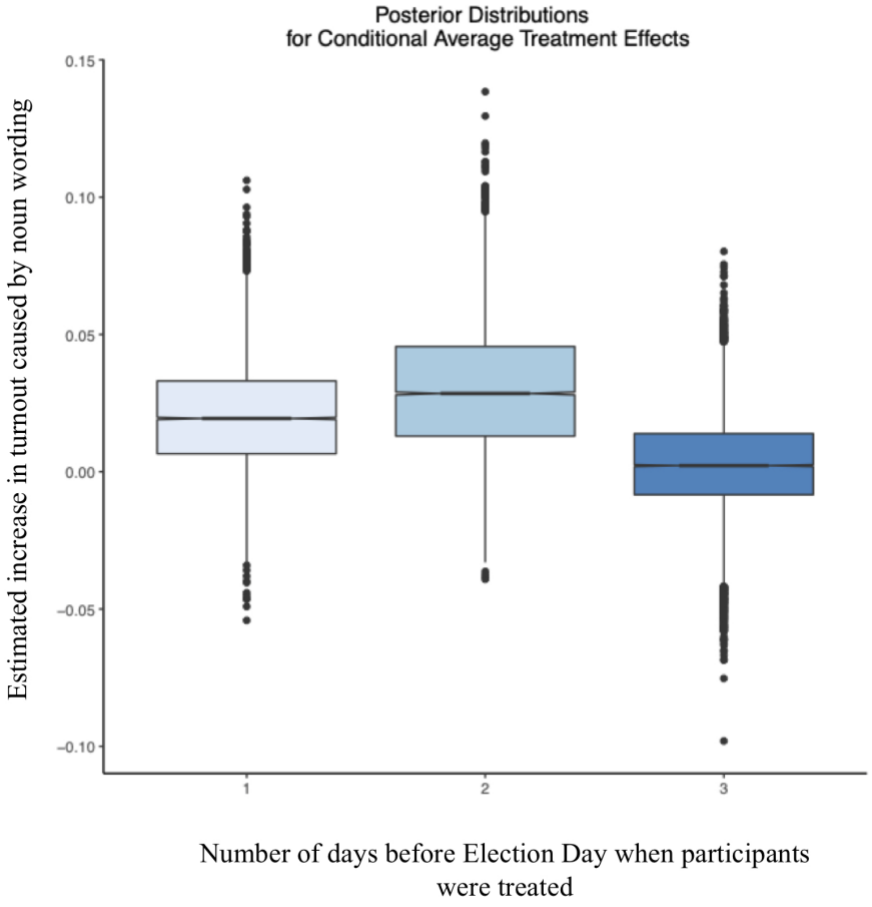

Figure S1. A Bayesian Causal Forest analysis shows that the noun-vs.-verb treatment effect on voter turnout is likely to be positive and greater than zero when participants complete the experimental manipulation 1-2 days before the election, but not 3 days before. Dots correspond to random draws from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