

How Experiments Help Campaigns Persuade Voters: Evidence from a Large Archive of Campaigns' Own Experiments*

LUKE HEWITT *Stanford University, United States*

DAVID BROOCKMAN *University of California, Berkeley, United States*

ALEXANDER COPPOCK *Yale University, United States*

BEN M. TAPPIN *Royal Holloway, University of London, United Kingdom*

JAMES SLEZAK *Swayable, United States*

VALERIE COFFMAN *Swayable, United States*

NATHANIEL LUBIN *Cornell Tech, United States, and Incite Studio, United States*

MOHAMMAD HAMIDIAN *Swayable, United States*

Political campaigns increasingly conduct experiments to learn how to persuade voters. Little research has considered the implications of this trend for elections or democracy. To probe these implications, we analyze a unique archive of 146 advertising experiments conducted by US campaigns in 2018 and 2020 using the platform Swayable. This archive includes 617 advertisements produced by 51 campaigns and tested with over 500,000 respondents. Importantly, we analyze the complete archive, avoiding publication bias. We find small but meaningful variation in the persuasive effects of advertisements. In addition, we find that common theories about what makes advertising persuasive have limited and context-dependent power to predict persuasiveness. These findings indicate that experiments can compound money's influence in elections: it is difficult to predict ex ante which ads persuade, experiments help campaigns do so, but the gains from these findings principally accrue to campaigns well-financed enough to deploy these ads at scale.


American political campaigns are among the most expensive in the world, and “television advertising is the cornerstone of many” of these campaigns (Sides, Vavreck, and Warshaw 2021,


715). For instance, Jacobson and Carson (2019) find that almost half of a typical congressional campaign's budget is spent on TV ads. The latest research finds that this advertising has small per-person persuasive effects that, across a large amount of advertising, can accumulate to meaningful effects on competitive election outcomes (Sides, Vavreck, and Warshaw 2021).


Given the enormous sums they spend on television advertising and that the main mechanism by which such advertising affects elections appears to be through persuasion not mobilization (Sides, Vavreck, and Warshaw 2021), a central challenge for campaigns is making this advertising maximally persuasive. Both campaign consultants and the academic literature are replete with theories of how to persuade voters in campaigns, especially through paid advertising. Empirically validating those theories is typically quite difficult because doing so requires estimating the causal effects of alternative advertisements. Campaigns and their consultants often use tools such as focus groups or intuition (Thurber and Nelson 2001) that, while potentially providing some important insights, may not reliably uncover the causal effects of competing persuasive strategies. Experimental work carried out by academics, too, has its flaws, as it often relies on treatments academics themselves compose or tests advertisements outside of a campaign context, undermining ecological validity.


* Luke Hewitt was previously employed by Swayable in May to November 2020.

Luke Hewitt , Senior Research Fellow, Polarization and Social Change Lab, Stanford University, lbh@stanford.edu.

Corresponding author: David Broockman , Associate Professor, Travers Department of Political Science, University of California, Berkeley, United States, dbroockman@berkeley.edu.

Alexander Coppock , Associate Professor on Term, Department of Political Science, Yale University, United States, alex.coppock@yale.edu.

Ben M. Tappin , Research Fellow, Department of Psychology, Royal Holloway, University of London, United Kingdom, benmtappin@googlemail.com.

James Slezak , Co-Founder, Swayable, United States, james@swayable.com.

Valerie Coffman, Co-Founder, Swayable, United States, valerie@swayable.com.

Nathaniel Lubin, Visiting Fellow, Digital Life Initiative, Cornell Tech, United States, and Founder and CEO, Incite Studio, United States, nate@incite.studio.

Mohammad Hamidian , Senior Experimental Scientist, Swayable, United States, mhh32@cornell.edu.

Received: June 21, 2022; revised: February 23, 2023; accepted: December 04, 2023.

Over the last decade, though, campaigns have increasingly conducted their own randomized experiments. These campaigns sometimes allow their experimental results to be shared publicly or in academic research (e.g., Kalla and Broockman 2018), but the vast majority remain the private information of campaigns or parties (Issenberg 2012). Moreover, even studies that are shared may be subject to publication bias (Franco, Malhotra, and Simonovits 2014). What lessons do these typically proprietary experiments hold for long-standing theories of what persuades voters? And what implications does the rise of campaign experimentation have for American politics?

In this article, we provide a rare window into these questions based on a unique collection of experiments campaigns themselves conducted in the midst of two election seasons in the United States (US). In 2018 and 2020, dozens of Democratic and other left-leaning campaign organizations contracted with the technology company Swayable to conduct survey experimental tests of their advertisements, intended to be later run on television and digitally. The extent of Swayable use underscores the prevalence of experimentation in modern American elections: Swayable was hired to conduct advertising experiments in support of Democratic candidates in 20 of the 36 (56%) US House races rated by the Cook Political Report as toss-ups prior to the 2022 election, and 100% of the US Senate races rated as toss-ups prior to the 2020 and 2022 elections (seven and four races, respectively).¹ Swayable's agreements with campaigns also allowed them to subsequently share the resulting data with academics for research.² The resulting set of experiments is an unprecedented treasure trove of data on the persuasive effects of campaigns' real ads, as tested in real time, during real campaigns, among voters eligible to vote in those elections, from randomized survey experiments. We executed an agreement with Swayable that allowed us to access the de-identified data from their entire universe of experiments, and that allowed us to publish whatever conclusions we drew from our analyses of those data.

To understand the implications of campaign experimentation, in this article, we analyze the complete archive of these experiments from the original microdata. We make small adjustments to account for Swayable's evolving design idiosyncrasies, described below. There are also some limitations to the data, such as incomplete information on attrition in some studies and the fact that these studies were originally conducted by a third party, described in greater detail below. We meta-analyze the resulting effect estimates to characterize average effects, what predicts these effects, and how variable these effects

are. With research assistance, we hand-coded theoretically important characteristics of each ad such as primary focus, messenger, emotion, tone, appeals, and use of evidence. We include these measurements as predictors in our meta-regression models of persuasive effects.

As we analyze the entire archive of these experiments, we can set aside concerns about publication bias; campaigns could not choose to cherry-pick which experiments we analyzed based on the results. By analyzing data from multiple election cycles and at multiple levels of office, we can also offer insights into how these conclusions vary across contexts.

Our investigation points to two main conclusions.

First, we find small but politically meaningful variation in ads' persuasive effects. Across the three contexts we examine (2018 downballot, 2020 downballot, and 2020 presidential), we find that the average ad affected immediately measured vote choice by 2.3 percentage points, 1.2 points, and 0.8 points, respectively. We estimate that the variance of the true treatment effects of the ads relative to these baselines is small but meaningful: our meta-analytic results indicate that the standard deviation of the distribution of true treatment effects were 1.5 points, 0.5 points, and 0.3 points, respectively, or roughly about half the size of the average effect. We characterize this scale of variation as "small" in absolute terms because even the largest of these amounts to just 0.03 standard deviations.³ But this variation is nevertheless politically meaningful, since it is common for advertisements to be 50% less or 50% more persuasive than the average advertisement. Although we expect the absolute size of these advertisements' effects to be much smaller in the field, when campaigns intend to deploy these advertisements to millions of people, we conduct simulations showing that choosing an above average ad over a below average ad could still easily determine the outcome of a close election. Our simulations also show that access to experimentation has profound implications for how campaigns should allocate their budgets.

Our second main conclusion is that the extant theories about what features of advertisements make them more effective have very limited and highly context-dependent explanatory power. We consider a variety of theories in the academic literature as well as those common among campaigns—for example, whether ads work better when they are negative, provide new facts about candidates, attempt to elicit emotions such as anger or enthusiasm, or feature testimonials. These predictions have attracted substantial attention over decades of research, and we offer one of the most comprehensive and systematic explorations of them to date. Assessing these predictions of when ads will be more or less effective across 617 real advertisements, we find at best limited evidence for any of them. Moreover, we find that "what works" changes from election to election. For example, features that predict stronger effects among ads produced for the 2018

¹ Swayable is one of several companies that provide this service. We do not have comparable figures for Swayable's competitors, so these figures represent a lower bound on the share of elections that feature advertisements tested in this manner.

² Data were shared under strict data security protocols negotiated between Swayable, MIT, UC Berkeley, and Yale in order to meet Swayable's client confidentiality obligations while permitting independent, replicable analysis, including anonymization of Swayable clients. Swayable also did not collect any personally identifiable information about respondents in the experiments.

³ The standard deviation of a binary variable is $\sqrt{p * (1-p)}$, so if the probability of voting for a candidate is 0.5, the standard deviation is 0.5 or 50 percentage points. $1.5 / 50 = 0.03$ standard units.

downballot elections fail to do so among ads produced for the 2020 downballot elections. Finally, even the theories that receive partial support explain a very small proportion of the overall variation in ads' persuasive effects. These findings suggest limits on the ability of general theories to predict the persuasiveness of any particular political advertisements across highly heterogeneous electoral contexts.

These two conclusions in turn have two implications for campaigns and for American politics more generally.

First, when it comes to campaign strategy, we conduct simulations that show that spending money on experimentation may be an extremely cost-effective investment. To the extent that ad experimentation does successfully identify more effective ads within an election cycle, modest investments to find the ads that work better would allow campaigns to dramatically increase the impact of their overall advertising spending. Indeed, our simulations find that the returns to ad experimentation may be so large that optimal campaign behavior would be to devote a substantial portion (over 10%) of their media budget to ad experimentation—a whole new category of expenditure that scarcely figures in classic theories of campaigning.

Second, we also demonstrate that this new behavior of campaigns has implications for American politics: *experiments increase the influence of money in elections*, because experimentation is a complement to campaign spending. In a world without experimentation, well-financed campaigns can of course “buy” more votes than less-well financed ones through more advertising (Sides, Vavreck, and Warshaw 2021). But experimentation may also serve as a multiplier that enhances the importance of financial advantages because it increases the marginal effects of advertising spending by making those ads more persuasive. Our findings thus suggest that scholars should not only view campaigns' use of experimentation as a way for researchers to generate knowledge about what persuades voters, but also as an object of study in itself with important implications for American elections and democracy.

Our article makes several contributions that advance our understanding of campaigns in an era of experimentation. First, we demonstrate theoretically and illustrate through simulations that the extent of heterogeneity in the effects of different campaign ads conditions the payoff of experimentation for campaigns: the more heterogeneous the effects of ads, the more campaigns gain from experimentation. Second, empirically, we estimate the extent of effect heterogeneity in campaign ads—in other words, how much campaigns have to gain from experimentation—using a unique archive of real ads tested by real campaigns during real elections. Appendix D of the Supplementary Material quantifies the extent of this empirical contribution, showing that without access to the large archive of experiments we analyze, neither scholars nor practitioners would be able to form reasonably precise estimates about how heterogeneous the effects of advertisements are and therefore how beneficial experiments would be. Third, we provide some of the most comprehensive tests available to date of influential theories of campaign persuasion, using real ads to test

them in the midst of real campaigns. We show that theories common among scholars or practitioners do not reliably predict ads' effects, suggesting that there may be no alternative to experimentation to determine which ads are most effective in a given electoral context. Fourth, we show that, given the extent and seeming unpredictability of heterogeneity in ads' effects, experiments are likely to increase the impact of money in elections, as they allow each dollar of campaign spending to persuade more voters. Finally, in the conclusion, we show how our theoretical analysis and empirical contributions suggest a series of priorities for future research about the promises and pitfalls of experimentation for campaigns.

WHAT WORKS TO PERSUADE VOTERS, AND CAN CAMPAIGNS LEARN WHAT DOES?

One approach for choosing how to craft persuasive appeals to voters is to rely on general theories of persuasion about which advertisements would be persuasive.

At first blush, we might have reasons to be optimistic about the potential for general theories to capture meaningful variation in the persuasiveness of advertisements: the closely related research literature on voter mobilization has uncovered a number of empirical regularities regarding which strategies successfully turn people out to vote across many experiments (Green, McGrath, and Aronow 2013).

Is it similarly possible to discern general patterns in what messages work to *persuade* voters? In this section, we review ideas about what features of ads make them more effective from two main sources: the academic literature on persuasion and the features that advertising practitioners highlight. Later in the article, we test whether general hypotheses about what features makes more ads persuasive are able to explain meaningful variation in ad effects or whether these hypotheses reach similar results in different contexts. As a result, as we review these hypotheses, we also describe the methodology we used for assembling them.

First, we systematically reviewed the academic literature on campaign persuasion with the goal of finding hypotheses that have drawn considerable attention from scholars. To do so, we used Google Scholar to search for studies of political ad persuasion, examined literature reviews and meta-analyses on the effects of campaign ads (e.g., Lau et al. 1999), and consulted with scholars in the field. This review largely focused on studies considering persuasion in US politics, although we considered many studies that were not necessarily country-specific (e.g., psychology studies that are often conducted in the US but advance theoretical claims not specifically limited to the US context). We then qualitatively evaluated on which hypotheses the literature collectively placed the greatest emphasis. This yielded several sets of hypotheses drawn from research in political science and the psychology of persuasion, which we review below and test in our analysis.

First, a voluminous academic literature considers the consequences for persuasion if advertisements are

produced with negative, contrast, or positive tone. For example, a meta-analysis based on a mix of 111 observational and experimental studies first presented in Lau et al. (1999) and updated in Lau, Sigelman, and Rovner (2007) finds no effect on vote choice on average (see also Ansolabehere et al. 1994).

Second, another major theme in the literature on campaign persuasion is source cues. As early as the Hovland studies of persuasion (e.g., Hovland, Janis, and Kelley 1953), academic theories have focused on the source of a message as a core feature (Iyengar and Valentino 2000; Weber, Dunaway, and Johnson 2012). Message source can serve as a group cue—if the receiver perceives the messenger to be a member of their in-group, then the message can serve as an information shortcut about which candidate the in-group supports. One salient in-group is “everyday people,” which is perhaps why many advertisements feature “average Americans” as the messenger. Another version of in-group persuaders of special interest in politics is messenger partisanship (Zaller 1992); out-partisan persuaders might be more effective in convincing out-partisan voters—for example, a Democratic campaign featuring Republicans who support a Democratic candidate. The consequences of the gender of the messenger have also received a fair amount of scholarly attention (e.g., Searles et al. 2020), although the review offered in Strach et al. (2015) uncovers little scholarly agreement on the effects of men versus women as messengers. Research on source cues generally also singles out experts on particular topics as potentially especially persuasive messengers (Iyengar and Valentino 2000).

A further strand of advertising theory considers whether the ad provides new information or facts (e.g., Broockman and Kalla 2022); a major theme of the research on campaigns, and on television ads in particular, is that they persuade in part by providing information to voters (for review, see Sides, Vavreck, and Warshaw 2021). These theories are often rooted in a spatial model of politics—the advertisement provides viewers with new information about the policies proposed by the candidate, so viewers who like that policy should increase their support for the candidate upon viewing the advertisement (e.g., Kendall, Nannicini, and Trebbi 2015; Vavreck 2001). In our empirical section, we distinguish between whether facts are about candidates’ background or about their issue positions and whether they are more or less specific.

Finally, a large literature in political psychology considers the mediating roles of emotion, most notably anger and enthusiasm, in the causal processes by which advertising affects outcomes (for review, see Albertson, Dun, and Gadarian 2020). For example, Brader (2005) finds that ads that stimulate enthusiasm boost political participation, while ads that stoke fear are more likely to affect vote choice.

We drew a second set of hypotheses from political practitioners. In particular, Swayable team members frequently meet with campaigners and their advisors, who raised and documented hypotheses that were shared with the company. The Swayable team reviewed

their notes and recollections of these meetings for hypotheses that campaigns frequently raised, many of which have parallels in the academic literature.

First, practitioners often noted that ads can vary in how “pushy” they are in making the case for a candidate—that is, how aggressive and explicit the ads are in instructing viewers what to do or think. Consistent with theories of psychological reactance (for review, see Bigsby and Wilson 2020), a possible concern is that voters will resist persuasion from ads that are too pushy but that voters will not follow the implications of ads that are insufficiently direct. Second, consistent with psychological theories of conclusion explicitness (e.g., O’Keefe 1997), some practitioners also focused on the specific ask (“call to action”) that the ad is making, asserting that candidate advertisements should include an explicit appeal to “vote for” the candidate. Next, practitioners sometimes wondered whether advertisements that look amateurish or low-budget would be less effective, giving rise to concerns about production value, consistent with theories of advertising as signaling quality to consumers (Nelson 1974). Finally, classic literature suggests that practitioners also rely on a series of particular rhetorical strategies: name calling, testimonials (Yourman 1939), metaphor (Schlesinger and Lau 2000; Thibodeau and Boroditsky 2011), and transfers of association are common techniques in the advertising toolkit.

Theoretical Argument

We began the previous section by noting that campaigns might rely on general theories of persuasion to help craft maximally persuasive advertisements. However, despite the large body of theoretical work predicting consistent patterns in what kinds of messages work to persuade voters, there are good reasons to expect that no such general patterns exist. Voters’ views and priorities shift over time, and persuasive messaging in real campaigns takes place within a competitive context where voters interpret one ad’s message in the context of other messages they have received (Vavreck 2009). For example, the effects of advertising a candidate’s popular issue positions versus their previous political experience may depend upon whether an opposing candidate has recently attacked the advertising candidate on either dimension, or on current events. As a consequence, even if certain ad features were to reliably predict persuasiveness in artificial settings (e.g., fake elections in a laboratory environment), or within a particular election, heterogeneity across real electoral contexts may make reliable generalization nearly impossible.

Such difficulties in predicting when persuasion will and will not work could help account for persuasion’s uneven track record (Kalla and Broockman 2018). After all, if it were easy to predict which persuasive interventions worked, why would campaigns run so many persuasion campaigns that appear to have minimal effects? Green and Gerber (2019) thus argue that “it is hard for a campaign to know in advance whether its persuasive message will resonate with voters” (182). Consistent with this line of reasoning, a recent study

(Coppock, Hill, and Vavreck 2020) found at best limited evidence of stronger effects associated with the messenger, sender, or tone of presidential advertisements. Furthermore, even if a sophisticated campaign is able to build up such knowledge within a given election, patterns in what is persuasive in one election may be context-specific and not generalize to other times or places (Munger 2019). Indeed, relatively little prior research has sought to examine whether features of persuasive communication identified as predictive of persuasive effects in individual pieces of research generalize across contexts, rather than just at one particular place and time (for important exceptions, see Blumenau and Lauderdale 2024; Vavreck 2009). Research in social psychology echoes this pessimism, noting that “persuasion phenomena are complicated, making the development of dependable generalization difficult” (O’Keefe 2004, 31).

This argument has important implications for elections and democracy. In particular, consider how campaigns should strategically respond if “dependable generalization” about what is persuasive is nearly impossible, and yet, *within* a given context at a given time, some messages may be more persuasive than others. One potential response is to devote resources to forms of research that can identify persuasive messages in a campaign’s particular context, even if they do not generalize beyond it. And indeed, campaigns regularly do so, often outsourcing this work to political consultants (Martin and Peskowitz 2015; Thurber and Nelson 2001).

Campaigns and their consultants have many techniques at their disposal for learning how to communicate with voters in any given election campaign. Traditional approaches include focus groups and face-to-face meetings, yielding qualitative insights into how voters react to different persuasive attempts (Thurber and Nelson 2001). Over the last decade, campaigns have increasingly turned to randomized trials to understand what messages and advertisements persuade more or less effectively (“experiments”) (Issenberg 2012). However, field experiments on television ads are still relatively scarce, likely because they are extremely expensive: Kalla and Broockman’s (2018) meta-analysis of field experiments contained only one published field experiment on candidate television ads (Gerber et al. 2011), and this experiment was still too small to compare alternative treatments. Given the central place of TV ads in modern elections, campaigns have therefore turned to randomized experiments conducted in survey contexts in order to determine which of their ads are most persuasive. Although comprehensive data on the extent of campaign experimentation by all companies and organizations are not publicly available, data from our partner Swayable illustrate this trend: Swayable tested ads in 20 of the 36 (56%) US House races the Cook Political Report rated as toss-ups prior to the 2022 election, up from only 6 of the 30 (20%) toss-up US House races in the 2018 election cycle.

Campaigns’ turn toward experimentation may have significant broader implications for campaigns and for

democracy. Insofar as these approaches are able to surface more effective advertisements, they will increase the impact of money in elections: the gains from finding the most persuasive ads principally accrue to campaigns well-financed enough to run these ads at scale. If different advertisements have very similar effects, experimentation has a limited payoff, as a campaign’s best ads would be not much more effective than their worst. However, if ad effects do vary, it can be advantageous for campaigns to spend substantial sums on experimentation in order to identify meaningfully more effective ads. As we show below, under the levels of heterogeneity we observe in the effects of real campaign advertisements, campaigns can substantially increase the returns to their advertising by testing a small number of ads with randomized experiments that are inexpensive, at least relative to the overall advertising budget. In the conclusion, we expand on our argument that this dynamic has subtle but important implications for both elections and American democracy. Because experimentation allows campaigns to increase the effectiveness of their advertising spending, experimentation increases the influence of money in elections, with benefits that redound principally to the best-funded campaigns.

As we show, though, these payoffs to campaigns and broader implications of the rise in campaign experimentation depend on the level of heterogeneity in the persuasive effects of different advertisements and how predictable this heterogeneity is in advance. Our unique data source allows us to shed light on these questions in unprecedented detail.

RESEARCH DESIGN

Experimental Design

We analyze data from 146 randomized survey experiments conducted by the technology platform Swayable during the 2018 and 2020 elections. All data in the 2018 studies were collected by Swayable between April 4, 2018 and March 15, 2019. All data in the 2020 studies (both downballot and presidential) were collected between December 7, 2019 and December 24, 2020.

Each experiment began when a campaign contracted with Swayable to measure the effectiveness of their potential advertisements. These clients included candidate campaigns, party coordinated campaigns, and independent expenditure campaigns. Campaigns informed Swayable of the relevant populations of subjects (the entire US or just select states) and uploaded multiple video advertisements they wished to test. As described previously, Swayable’s agreements with campaigns allowed them to share de-identified data from these experiments with academics for research purposes. We executed an agreement with Swayable that allowed us to access these de-identified data and publish whatever conclusions we reached.

Our study is restricted to “candidate persuasion” videos—that is, videos which aim to increase electoral support for a particular candidate (always a Democrat)

and/or reduce electoral support for a particular candidate (always a Republican)—and we therefore exclude any ads which instead aim to persuade viewers about issues, or purely to increase turnout. We also exclude treatments which were simply static images, text, or which did not contain audio. Most of the experiments are from general elections, although the 2018 data in particular feature some primary elections.

Importantly, our dataset contains every Swayable experiment meeting these criteria, and our data availability is not conditioned on the results of any of the experiments. This feature of the dataset is important, as individual experiments conducted by or in collaboration with campaigns are often subject to publication bias, either because campaigns might not agree to release null or positive findings (to avoid embarrassment or, alternatively, to avoid sharing successful tactics with the opposition), or because researchers may not bother to publish null results (Franco, Malhotra, and Simonovits 2014). Our analysis of Swayable studies is in a rare category of meta-analysis where we can be confident we are analyzing the full universe of conducted studies.

Swayable recruited subjects online through their proprietary acquisition channels. In line with our pre-analysis plans (described below), we exclude responses flagged as spammers and as duplicate responses. In a deviation from the PAPs, we also exclude two further sets of subjects: “mistargeted” respondents (e.g., the client sought respondents from a specific state only, but some out-of-state respondents slipped through) and “inattentive” subjects as identified by a pre-treatment attention check. These data quality measures are consistent with the measures used by Swayable in analysis presented to campaigns.

Once in the survey environment, subjects were asked pre-treatment demographic questions. The precise set that were asked differs somewhat from study to study. When these pre-treatment questions were asked, they were in a forced-response format, meaning that the covariate data exhibit no item-nonresponse. Where available, we include pre-treatment measures of subjects’ age, gender, race/ethnicity, education, party identification, liberal-conservative ideology, and Trump job approval. The precise wording of these questions and their distributions are presented in the Supplementary Material. For the 2018 studies, these questions were asked post-treatment, and so we do not include them as covariates in our analysis.

Subjects were then randomized into treatment conditions or a placebo. The placebo was a public service announcement video on a non-electoral issue that differed between studies—for example, smoking or texting while driving.

One complication that arises in these data is that Swayable’s treatment assignment scheme was somewhat nonstandard. Because they wished to have more evenly sized treatment arms, in the 2018 data, treatments were cluster-assigned based on the time at which respondents entered the survey. At the start of each cluster (lasting 1–4 minutes), the target sample size for each treatment condition was defined as the ideal

number of responses to that treatment given the current sample size. The cluster was then assigned to the treatment with the fewest responses relative to its target. To handle this complication, we cluster our standard errors at the level of these clusters in our analysis in the 2018 data. In the 2020 data, treatments were assigned randomly at the individual level, but using dynamically updating assignment probabilities. For each respondent, each treatment was sampled in proportion to the remaining number of responses it would need before reaching its target. Unfortunately, these assignment probabilities were not saved, but data on covariate balance suggest that there is good balance on both baseline demographic and political covariates, and better-than-expected balance on time, indicating that this assignment scheme functioned essentially as implicitly blocking on time. The consequences of this procedure for sampling variability are, therefore, likely trivial.

A second complication is that, in a small number of cases (four studies in 2018, seven studies in 2020), Swayable either added or subtracted a treatment at some point during the experiment, then used an algorithm to dynamically update what fraction of subjects were assigned to treatments over time. This procedure generates different probabilities of assignment for different units that are difficult to reconstruct. Our solution is to simply discard any data collected after the moment of addition or subtraction of the new treatment (detected as a treatment with no responses collected in the initial or final 10% of the study). This procedure was not declared in the PAP.

A third complication and limitation of the data arises from the fact that, in most of the surveys, Swayable did not keep records of data for people who were assigned a treatment but did not provide outcome data, so we cannot directly verify whether the treatments induce differential attrition, which would be a threat to inference. There are two versions of this problem: in 2018, individuals could skip questions (on which we do have records) or leave the survey (on which we do not); in 2020, individuals could not skip questions but some still left the survey (on which we have records for some studies). In short, we find evidence of differential attrition driven by skipping survey questions in the 2018 data, largely driven by the placebo group. In the 2018 data, differential attrition appears modest between treatment arms, so our estimates using study fixed effects are plausibly unaffected. In the 2020 data, we see limited evidence of differential attrition. As detailed in Appendix B.3 of the Supplementary Material, we probe the problems potentially posed by differential attrition in four ways. First, we demonstrate balance on covariates by treatment condition, among those units who complete the survey. These balance checks are presented in Appendix A.3 of the Supplementary Material. Overall, we find good covariate balance in both the 2018 and 2020 data. Importantly, this includes good balance on baseline political attitudes, not only demographic attributes. Second, for a subset of the surveys in 2020 ($N = 22$), Swayable’s engineers were able to reconstruct a dataset of all subjects who ever started the survey. The average

TABLE 1. Summary of All Three Datasets

| | First study | Last study | Total N | # Treatments | N per treatment | Vote choice only | Favorability only | Both outcomes |
|-------------------|-------------|------------|---------|--------------|-----------------|------------------|-------------------|---------------|
| 2018 | 2018-04-09 | 2019-03-15 | 93,969 | 137 | 479 | 3% | 33% | 63% |
| 2020 downballot | 2019-12-07 | 2020-12-22 | 101,782 | 189 | 348 | 27% | 5% | 67% |
| 2020 presidential | 2020-02-13 | 2020-10-24 | 302,589 | 292 | 717 | 39% | 0% | 60% |

Note: Full study-level tables are available in Dataverse Appendix DA1.

rate of post-treatment attrition was approximately 4%, and in only one of these 22 studies, do we find a statistically significant effect of treatment on survey completion. Third, we further demonstrate that any effects on response within these studies do not differ by covariates—that is, we do not find evidence that a particular kind of subject (e.g., Republican subjects) are systematically more or less likely to drop out in the treatment conditions versus the placebo condition. Fourth and lastly, in Appendix B.3 of the Supplementary Material, we present an analysis of this subset of studies that compares our unweighted treatment effect estimates to inverse-probability-weighted estimates (using the procedure described in Gerber and Green 2012, chap. 7), finding virtually no differences in estimated effects. We come away from these four design checks assured that differential attrition is not a serious source of error in our measurement of the persuasive effects of these advertisements. See Appendix B.3 for further discussion.

A final complication is that, for a subset of the studies (2018 only), certain post-treatment outcome questions included a “no opinion” response option. Following our pre-analysis plan, these missing outcomes are simply excluded from analysis; however, this approach may induce bias because, unlike the case of overall attrition (which we found to be independent of treatment), respondents were significantly more likely to select “no opinion” when in the placebo group than the treatment group. Our best guess is that this pattern of “no opinion” response is due to a lack of familiarity with the candidates among placebo subjects. We do not find evidence of different rates of “no opinion” responses between the treatment groups in the same experiment, nor interactions with demographic covariates. This pattern suggests that dropping “no opinion” answers results in a constant shift in the estimated treatment effects within each study, but should not impact our estimate of the variability of treatment effects within studies. In Supplementary Figure OA12, we consider a range of alternative approaches to handle these missing outcomes—including inverse probability weighting, as well as fitting meta-regressions with study-level intercepts—and find that our results remain fairly stable regardless of the choices made.

Outcome Measures

We focus on two post-treatment outcome measures, vote choice and favorability, both initially measured on

a 0–10 Likert scale (e.g., “How likely are you to vote for Donald Trump in the 2020 presidential election?” or “How favorable do you feel towards Donald Trump?”).⁴ We rescale all variables to range from 0 to 100 to aid interpretation. Where a survey contains multiple vote choice (or favorability) questions, such as support for the candidate and for their opponent, we create a vote choice (or favorability) index as weighted average of these questions.⁵ Swayable itself uses a pre-defined primary index, which was the outcome as specified in the 2018 PAP. In an unregistered robustness check (see Appendix B.2 of the Supplementary Material), we also dichotomize the vote choice scale, recoding it so that values of 0–4 are equal to 0, 5 is equal to 0.5, and 6–10 is equal to 1, allowing us to compute an estimated effect on vote share.

For each of the three sets of experiments (2018, 2020 downballot, and 2020 presidential), Table 1 shows the date ranges, total number of subjects, total number of treatments tested, and what fractions of them measure vote choice and favorability. Detailed summary tables for each separate study are available in Dataverse Appendix Table DA1.

Measuring Advertisement Characteristics

Some of our hypotheses concern how the treatment effects of ads vary with features of the advertisements themselves, such as their tone or whether they used particular persuasive techniques (see the previous section for review). We measured the characteristics of advertisements using human coders in two rounds. The 2018 advertisements were tagged by eight undergraduate research assistants, with three independent ratings per video. The 2020 videos were tagged by 11 different assistants, with two independent ratings per video. For each item, we first apply a simple adjustment for rater-biases, shifting the ratings such that all raters have equal

⁴ In all cases, the intention of this advertising was to influence vote choice. Treatment effects were typically larger on measures of candidate favorability.

⁵ If there is only a vote choice question for the Democratic candidate, this index is simply that question. If there is only a vote choice question for the opposition candidate, this index simply flips the sign. If both questions were asked, the index is the average of the two. If there are multiple opposition candidates (a very rare occurrence), the index first averages the opposition candidate items together before reversing the sign and averaging it with the Democratic candidate.

mean ratings. We then aggregate across raters by taking a mean of the multiple ratings for a single item. This correction was introduced in the 2020 PAP, in a deviation from 2018 PAP. An important feature of the ads is the election type—for example, Senate versus presidential election. For any videos where the research assistants disagreed about election type, we manually labeled these using the names of the candidates found in the outcome questions.

Inter-rater reliability varied substantially between items. Research assistants were highly consistent in their ratings of the most explicit features such as the race and gender of the primary messenger (single-rater ICC > 0.8), but less consistent in their ratings of the most subjective characteristics, such as how “pushy” the ad was (single-rater ICC = 0.23). Overall, single-rater reliability was higher in 2020 compared to 2018, although this is offset by the number of ratings per video. In Appendix A.4 of the Supplementary Material, we provide reliability estimates for each individual item and demonstrate that our findings hold within high-, medium-, and low-reliability items, suggesting that our conclusions are not a byproduct of low-reliability items. So that readers can gain a sense of the distribution of these features, we report a series of descriptive analyses (histograms, correlations, and time trends) in Dataverse Appendix DA2.1.

Analysis Procedures

Our analysis takes place in two steps: first, we estimate the treatment effects of each ad; and second, we meta-analyze these estimated treatment effects to make inferences about the unobserved true treatment effects of the advertisements.

First, we estimate the effects of each ad with ordinary least squares regressions of the outcome on indicators for each treatment and the pre-treatment covariates mentioned above. We include the covariates in order to obtain a more precise estimate of the treatment effect (Gerber and Green 2012, chap. 4). As registered, regressions in the 2020 data use HC2 robust standard errors; regressions in the 2018 data use CR2 clustered standard errors for reasons described above.

Second, we conduct random-effects meta-analyses using the resulting treatment effect estimates to make inferences about the distributions of unobserved true treatment effects of the ads. Importantly, these meta-analyses all take into account the sampling variability (i.e., standard errors) associated with the estimated treatment effects of each ad. They also take into account the fact that the estimates within a given experiment are correlated because all treatment groups were compared to a common placebo condition (Borenstein et al. 2021, chap. 30).⁶

⁶ In particular, instead of relying only on the ad-level estimates and standard errors, our meta-analytic estimators at the ad-level use a block-diagonal variance–covariance matrix, where the blocks are the (robust) variance–covariance matrices from each study. The use of this more conservative procedure is a small deviation from the 2018 PAP.

We conduct two forms of meta-analysis. First, we analyze the variation in the treatment effect *estimates* to make inferences about how much variation there is in the *true* treatment effects of each ad; meta-analyses are able to make such inferences by decomposing variation in the effect estimates into sampling variability, which is quantifiable, and variation in the true underlying effects, which accounts for remaining variation. This analysis allows us to make conclusions about the extent to which the true effects of ads vary between ads, even though we cannot observe the true effects of any of the ads in our sample. For instance, are all ads similarly effective—something which might be true even if the estimated treatment effects vary due to sampling variability? Or, do the true effects of ads meaningfully vary?

Second, we use meta-regressions to estimate whether certain features of ads are correlated with persuasiveness. Individual ads’ treatment effect estimates and their associated standard errors constitute the observations in these regressions; the meta-regressions then test hypotheses about whether ads with certain features are more persuasive than others, taking into account the statistical uncertainty associated with our estimates of the effects of each ad.

We present two basic sets of meta-regression analyses—those that can be conducted on both the 2018 and 2020 data (such as Positive vs. Negative tone), and those for which the characteristics were only labeled in the 2020 data (such as the political party of the messenger). For our 2020 PAP, we pre-registered our primary tests as these “new” characteristics, plus replications of any characteristics that were found to be significant in the 2018 data ($p < 0.05$). However, when presenting 2018 and 2020 results together, we group our analyses based on whether they were initially considered to be primary hypotheses in the 2018 PAP, to avoid selecting on the dependent variable.

To test each hypothesis, we run a separate meta-regression of the persuasion estimates on the relevant set of advertisement features. We estimate each meta-regression separately for 2018 downballot races, 2020 downballot races, and 2020 presidential races. We also estimate meta-regressions for vote choice and favorability separately. All downballot meta-regressions include indicator variables for race type, with a separate category for “Georgia US Senate Runoff.” In a robustness check reported in Table 2, we study these same models with the inclusion of election and study fixed-effects.

We average the meta-regression results across election type to obtain an overall estimate for each research question. In order to assess how the estimates change across election type, we likewise take the differences in the meta-regression estimates across election types.

RESULTS

The Distributions of True Persuasive Effects

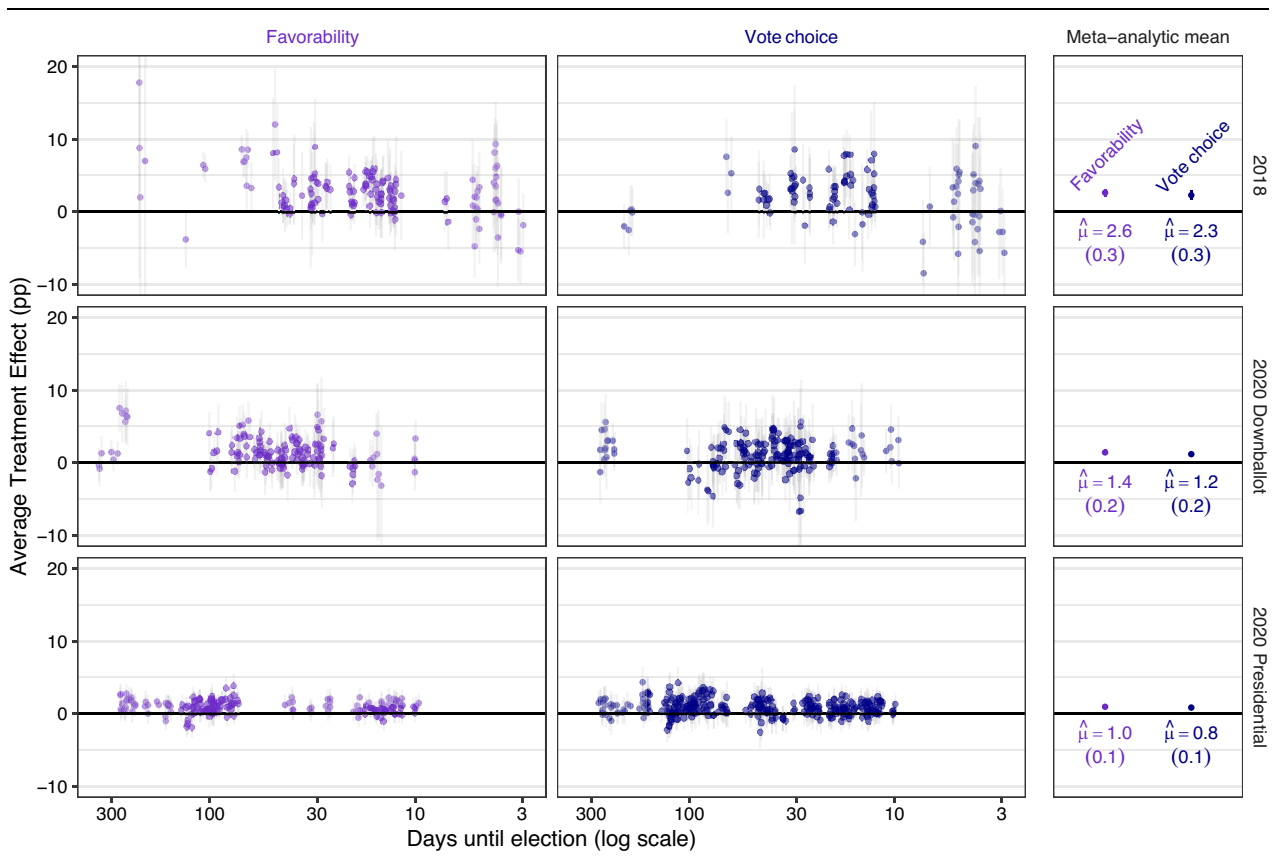
Figure 1 plots the average treatment effect *estimates* on the vertical axis and the days until the election on the

TABLE 2. Tests for (Residual) Heterogeneity

| Outcome | Election | No moderators | | | Race fixed effects | | Study fixed effects | |
|--------------|-------------------|----------------------|----------------------|------------|----------------------|------------|----------------------|------------|
| | | μ | τ | p -value | τ | p -value | τ | p -value |
| Vote choice | 2018 | 2.27 [1.60, 2.94] | 1.42 [0.90, 1.99] | < 0.001 | 1.43 [0.92, 2.01] | < 0.001 | 1.26 [0.75, 1.82] | < 0.001 |
| Vote choice | 2020 downballot | 1.18 [0.86, 1.51] | 0.47 [0.00, 0.85] | 0.054 | 0.42 [0.00, 0.81] | 0.128 | 0.37 [0.00, 0.78] | 0.247 |
| Vote choice | 2020 presidential | 0.85 [0.69, 1.00] | 0.34 [0.14, 0.50] | 0.001 | 0.34 [0.14, 0.50] | 0.001 | 0.33 [0.10, 0.48] | 0.011 |
| Favorability | 2018 | 2.62 [2.10, 3.14] | 1.67 [1.29, 2.11] | < 0.001 | 1.42 [1.04, 1.87] | < 0.001 | 1.07 [0.72, 1.46] | < 0.001 |
| Favorability | 2020 downballot | 1.42 [1.07, 1.77] | 0.85 [0.50, 1.20] | < 0.001 | 0.83 [0.49, 1.18] | < 0.001 | 0.74 [0.38, 1.08] | 0.001 |
| Favorability | 2020 presidential | 0.96 [0.79, 1.12] | 0.44 [0.29, 0.58] | < 0.001 | 0.44 [0.29, 0.58] | < 0.001 | 0.42 [0.26, 0.56] | < 0.001 |

Note: Each row in the table shows a set of a results for a given outcome in a given electoral context. The μ term shows our estimate and the associated 95% confidence interval for the average treatment effect of ads on that outcome in that context. The τ term gives our estimate of the standard deviation of the true treatment effects. The p -value is from a Q -test testing the null hypothesis that the true underlying treatment effects are homogeneous. The estimates under the Race fixed effects and Study fixed effects headings show estimates pertaining to the residual standard deviations after accounting for race- or study-fixed effects, respectively. The full models for the coefficients in this table are available in Dataverse Appendix Tables DA4–DA6.

FIGURE 1. Treatment Effect Estimates by Outcome and Time to Election



Note: Left: Unpooled treatment effect estimates on vote choice and candidate favorability, arranged chronologically by date of study. Within each column, each point shows the ATE for a unique treatment, with 95% confidence intervals. Right: Meta-analytic estimate of mean across all treatments ($\hat{\mu}$), with standard errors. 95% confidence intervals are plotted but are too narrow to be visible. For full model specifications, see Dataverse Appendix Table DA4.

horizontal axis on a log scale, faceted by outcome variable and election type. These plots give a sense of the raw data. Consistent with Kalla and Broockman (2018), we find some evidence that the effects generally decline as the election draws closer; across the three contexts, we find meaningful decreases in effects on favorability closer to election day, although this pattern does not replicate for vote choice (see Supplementary Table OA1).

Our first goal is to estimate the properties of the distribution of the *true* treatment effects of the ads, in particular its mean (μ) and standard deviation (τ). Random-effects meta-analysis provides estimates of both, appropriately accounting for the uncertainty in the effect estimates. To be abundantly clear, this analysis estimates the standard deviation of the *true* treatment effects (what Borenstein et al. 2021, 106, call T), and does *not* simply report the observed standard deviation of the estimated effects.

Table 2 reports the results of this procedure. In particular, in 2018, the average estimated effect on immediately measured vote choice of a single ad (μ) was 2.3 percentage points; in 2020 downballot races, the estimated effect was halved to 1.2 percentage points. In the 2020 presidential races, the average estimated effect was smaller at 0.8 percentage points. Relative to these point estimates for the average of the treatment effects, the estimates for the standard deviations of the treatment effects are *small but meaningful*. In 2018, the estimated standard deviation of the true treatment effects (τ) was 1.5 points, but in both the downballot and presidential races in 2020, the standard deviations were smaller, at 0.5 and 0.3 percentage points, respectively.⁷ These standard deviations are small in absolute magnitude, as are the effects of the ads themselves. However, the ratio of these figures, $\frac{\tau}{\mu}$, is 0.51 on average, indicating intuitively that it is commonplace for ads to be 51% more or less effective than the average ad. Such differences are nevertheless meaningful for campaigns for whom a 50% increase in effectiveness, multiplied by millions of voters, would yield a meaningful increase in votes.

Figure 2 overlays the random-effects estimated distributions of true treatment effects on top of histograms of the estimated treatment effects of each individual ad. Effects in 2020 are more tightly clustered around a lower mean compared with 2018, where effects are more dispersed around a higher mean.

We conduct several robustness checks on these findings. First, as reported in Appendix B2 of the Supplementary Material, the estimates are similar when we analyze a dichotomized version of the vote share variable. Second, the rightmost columns of Table 2 also report that results are similar when including experiment fixed effects to focus on within-experiment variation in effects only, indicating

that our conclusions are not driven by variation in how persuasive ads are across contexts or by differences in how persuadable different samples of participants were in different experiments (since the samples are constant within individual experiments). Third, Dataverse Appendix DA4 shows the results of a simulation indicating our meta-analytic procedure is able to distinguish between the presence of sampling variability and true variation in underlying treatment effects; in particular, in simulated data generated under the null where the true treatment effects of ads are identical (but the estimated effects vary due to sampling variability), our meta-analytic procedure only yields p -values less than 0.05 on a Q -test for heterogeneity approximately the expected 5% of the time.

Do Features of Advertisements Predict Their Persuasiveness?

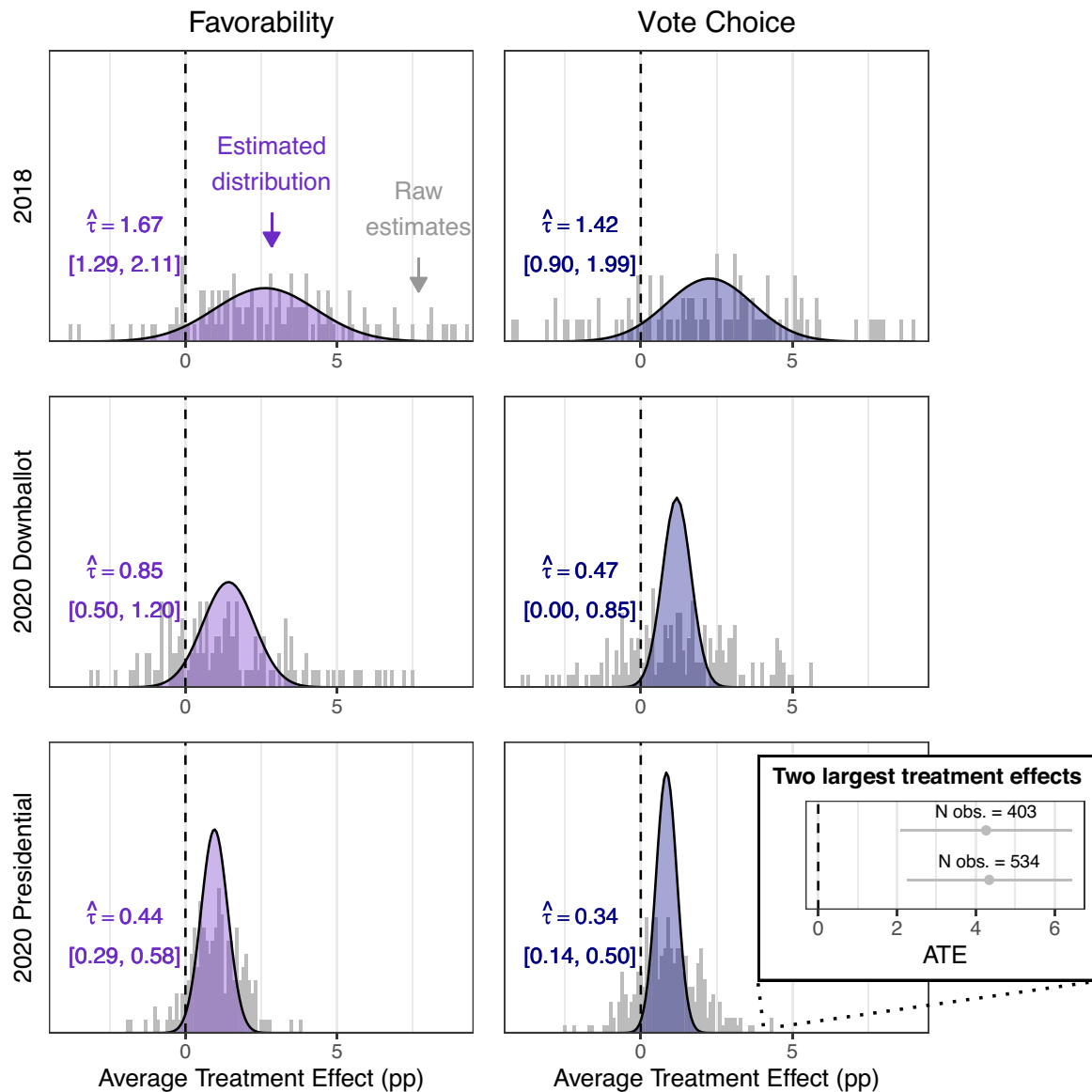
Thus far, we have described the variation in the effects of advertisements, but left to answer is whether one can predict which ads perform best from features of advertisements that academic theorists and political practitioners emphasize as important. In this section, we report the results of meta-regressions that assess the association of ad-level characteristics and the ads' effectiveness. We stress that these are descriptive contrasts: the characteristics of these advertisements were not randomly assigned, nor are the ads guaranteed to be otherwise similar but for the characteristics included in each regression.

Figure 3 provides an overview of our results. Each cell in Figure 3 reports the t -statistic from our meta-regressions (in Supplementary Figures OA2–OA7, we show the point estimates and standard errors for these estimates, but this matrix presentation allows us to compactly compare across contexts). The two large columns separate out the results on favorability and vote choice. Within these two large columns, we show three smaller columns separating out the results in 2018, the 2020 downballot elections, and the 2020 presidential elections. Each of the rows shows the hypotheses we tested, grouped by whether the 2018 PAP categorized them as primary or secondary hypotheses, then with a group of new hypotheses in the 2020 PAP below.

The unifying theme of the results is inconsistency. Across the 39 opportunities, we observe just one case in which the coefficient estimates across election types were all statistically significant and had the same sign: In both the 2020 downballot and 2020 presidential, “how pushy” the ad was a significant and positive predictor of effectiveness—but only for the favorability outcome, not the vote choice outcome. In the remaining 38 opportunities, either the sign or the significance of the contrast varied across contexts.

Another way of quantifying how weakly these features predict effects on vote choice and favorability is to estimate R^2 statistics for each meta-regression. The tables in Appendix A.4 of the Supplementary Material present these statistics. When we include all primary

⁷ As reported in Table 2, Q -tests indicate that we can conclusively reject the null hypothesis that the true treatment effects are homogeneous in five of six cases ($p \leq 0.001$), while the p -value is 0.054 in the sixth case.

FIGURE 2. Estimated Distribution of ATEs in Each Set of Experiments, after Accounting for Measurement Noise

Note: The figure shows the estimated distribution of true average treatment effects in each set of experiments (“Estimated distribution”), after accounting for sampling variability. The estimated treatment effects of each individual ad (“Raw estimates”) are plotted in gray. The meta-analytic estimate for the standard deviation in ATEs ($\hat{\tau}$) is given with 95% confidence intervals. For full model specifications, see Dataverse Appendix Table DA4.

hypotheses in one meta-regression, the adjusted R^2 values range between 0.12 and 0.34 for favorability and between 0 and 0.32 for vote choice. These statistics indicate that even when features do significantly correlate with treatment effects in a specific electoral context, they do not explain a meaningful share of the overall variation in effects.

Dataverse Appendix DA5 provides versions of Figures 2 and 3 by respondent gender and partisanship, finding that our results are broadly similar when restricting the sample to gender and partisan subgroups

(which is reassuring given the slight overrepresentation of women in the sample, as is common in online convenience samples (e.g., Berinsky, Huber, and Lenz 2012, Table 2)). In particular, we find that heterogeneity in ad effects is small but meaningful among all subgroups, and correlations with ad features remain inconsistent across electoral contexts. That said, we do find some evidence that average persuasive effects are somewhat larger for political independents than for partisans and we find that average effects are similar for men and for women.

FIGURE 3. *t*-Statistics for All Pre-Registered Meta-Regressions

| | Favorability | | | Vote choice | | | |
|----------------------------------|---|-----------------|-------------------|-------------|-----------------|-------------------|--------|
| | 2018 | 2020 Downballot | 2020 Presidential | 2018 | 2020 Downballot | 2020 Presidential | |
| 2018 Primary hypotheses | Candidate facts | 2.61* | 2.65** | -0.51 | 1.54 | 2.43* | -0.91 |
| | New fact (where fact present) | -1.88 | 2.29* | 0.55 | -1.29 | 0.73 | -0.35 |
| | Policy facts | -0.49 | -0.03 | 1.20 | 0.20 | -0.44 | 2.39* |
| | Primary focus: Candidate | 2.63** | 0.86 | -0.57 | 0.48 | 0.12 | -1.02 |
| | Primary focus: Issues | 3.46** | -1.60 | 2.75** | 2.57* | -1.12 | -0.55 |
| | Technique: Negative name-calling | 0.84 | 0.85 | -0.02 | 0.81 | -0.60 | 0.64 |
| | Technique: Negative testimonial | 1.57 | 1.39 | -0.25 | 1.24 | 1.82 | -0.29 |
| | Technique: Negative transfer of association | 0.13 | 0.88 | -0.16 | -0.20 | 0.11 | 0.50 |
| | Technique: Plain folks | -0.26 | 0.29 | 1.10 | 1.42 | 2.03* | 1.41 |
| | Technique: Positive name-calling | 0.30 | -0.95 | -0.91 | -0.73 | -0.63 | 0.28 |
| | Technique: Positive testimonial | 0.75 | -0.71 | 1.11 | -1.34 | -0.64 | 1.04 |
| | Technique: Positive transfer of association | -0.31 | -0.75 | 0.14 | 0.83 | 2.18* | 0.27 |
| 2018 Secondary hypotheses | Cited fact (where fact present) | -2.07* | 1.34 | -0.54 | -0.82 | 0.25 | -1.28 |
| | Emotion: Anger | 0.96 | 2.16* | -0.64 | 3.02** | 1.08 | 0.93 |
| | Emotion: Enthusiasm | 1.21 | -1.68 | -0.38 | -0.27 | 0.09 | 0.74 |
| | Explicit vote for | 2.01* | -2.75** | 1.59 | 2.83** | 0.47 | 1.82 |
| | Messenger: Female | 0.30 | 0.70 | 2.66** | 1.79 | -1.21 | 2.27* |
| | Messenger: Politician | 2.58* | 1.21 | -1.64 | -0.18 | 1.36 | -0.39 |
| | Primary tone: Contrast | 0.52 | -0.53 | 0.72 | 0.58 | -1.20 | 3.75** |
| | Primary tone: Positive | 0.14 | -2.71** | 0.17 | -1.47 | 0.60 | 1.89 |
| | Production value: High | 1.03 | 2.58* | -0.29 | 0.50 | 2.03* | -0.17 |
| | Specificity: Candidate facts | -0.12 | 3.45** | -1.26 | 1.32 | 0.92 | -1.79 |
| | Specificity: Policy facts | -0.33 | 1.30 | 2.38* | -0.20 | -0.12 | 1.29 |
| | 2020 New hypotheses | How pushy | | 2.74** | 2.24* | | 0.85 |
| Issue: BLM/Race | | | 0.02 | -1.84 | | 0.18 | -0.62 |
| Issue: COVID-19 | | | -2.35* | 1.12 | | -0.20 | -0.46 |
| Issue: Decency | | | 3.78** | -2.23* | | 1.38 | 0.54 |
| Messenger: Everyday people | | | 0.96 | 1.63 | | -0.39 | 0.86 |
| Messenger: Healthcare worker | | | -1.48 | 0.46 | | 0.03 | 1.33 |
| Messenger: Republican | | | 1.91 | -0.37 | | 2.62** | 1.18 |

Note: Each row corresponds with one hypothesis and each column corresponds with one dataset. The cells record the *t*-statistics on the meta-regressions testing each hypothesis in each dataset, which also maps to the cell colors, which range from purple (most positive values), to white (zero), to orange (most negative values). Full model specifications are provided in Dataverse Appendix DA2.

THE RETURNS TO EXPERIMENTATION AND THE IMPLICATIONS FOR ELECTORAL POLITICS

Taken together, our results point to two conclusions: first, the variation in the treatment effects of political advertisements is small but meaningful. In an absolute sense, the differences in effectiveness are small, but the relative differences between these effects are large enough that campaigns could meaningfully benefit from running an ad that is a standard deviation more effective than the average ad. Second, predicting which of a set of possible ads is the most effective is very difficult, since “what works” changes with context. In other words, some ads are better than others, but it is difficult to predict which ones those are in advance; campaigns can, however, turn to experiments to help identify the most effective advertisements.

These conclusions have important implications for elections and democracy. A subtle but important consequence we explore in this section is that campaign experimentation has the potential to increase the impact of money in politics. In particular, in this section, we show that experimentation is not only a good investment, but that it also increases the return to campaign spending because it makes each dollar of a campaign’s budget go further. In other words, experimentation enhances the importance of financial advantages because it increases the marginal effects of advertising spending.

We illustrate these implications using three simulations, motivated by the example of a typical US Senate campaign. Importantly, our simulations account for the fact that the survey experimental estimates from Swayable’s studies are likely to overestimate the effects of advertisements in the field (due to differences in ad delivery format, decay over time, etc.): we anchor our simulations to the estimated cost per vote of TV campaign advertising of \$200 per vote found in prior work (Sides, Vavreck, and Warshaw 2021) (for further discussion, see Appendix C1 of the Supplementary Material). We then use the survey experimental results to estimate the spread of effects around this mean. The key assumption of this approach and our simulations is that both the average effects of ads and the variability of their effects scale down proportionally from the survey environment to the field. We discuss this assumption in further detail in the conclusion.

Our first simulation considers how much money it would be optimal for campaigns to spend on

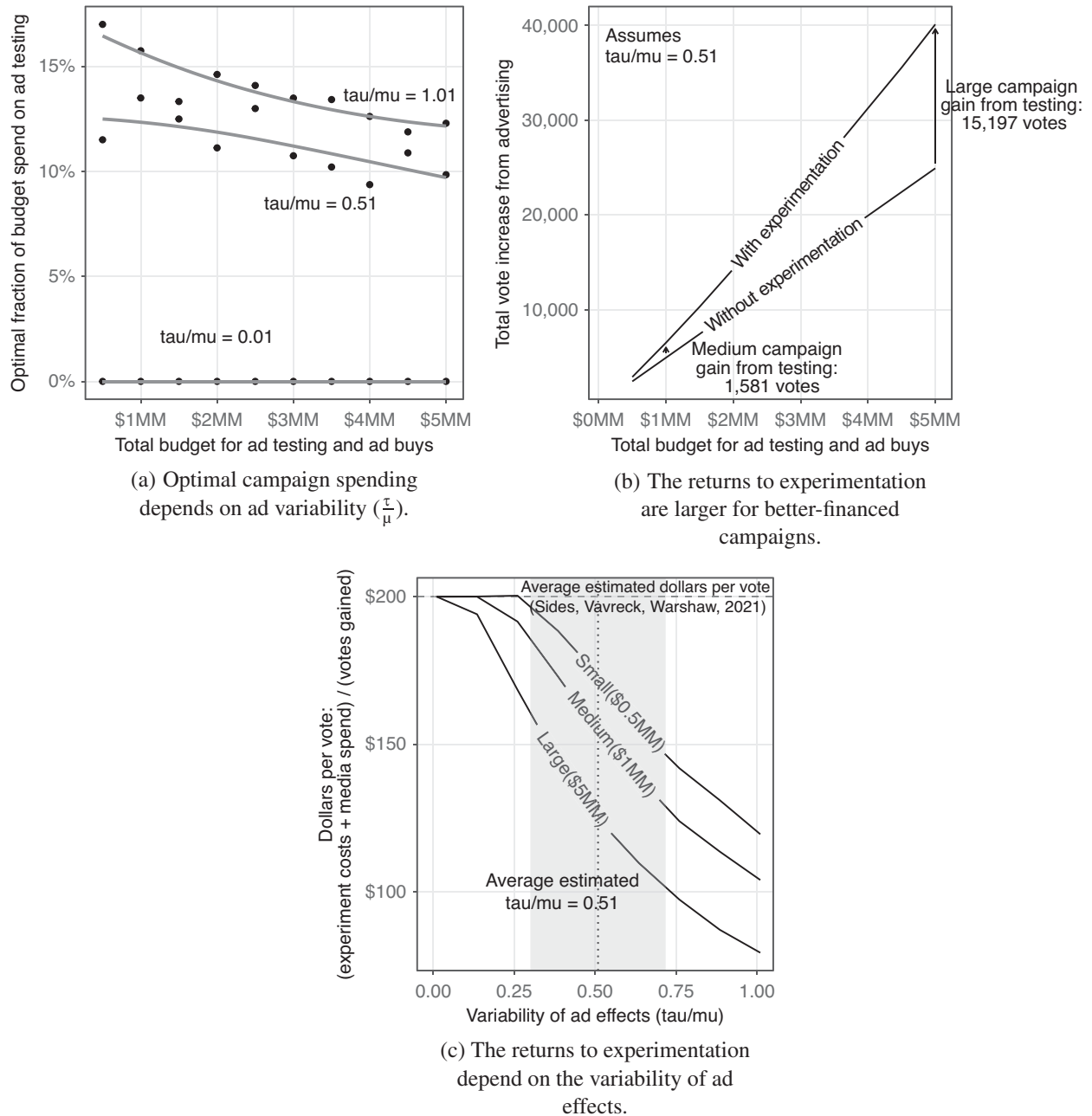
experiments given the trade-off they face between investing money in experimentation to make their ads more persuasive and simply using that money to air ads. In this simulation, we consider campaigns with media budgets ranging from \$500,000 to \$5,000,000. At each budget level, we consider how campaigns should allocate their funds if they believed ad variability ($\frac{\epsilon}{\mu}$) were tiny (0.01), of the magnitude we estimate (0.51), or large (1.01). To do so, the simulations consider a range of scenarios for the number of ads campaigns could produce to test and the number of experimental subjects they could recruit to their experiments; these two costs together represent the cost of ad experimentation. To determine how much money should be invested in ad experimentation, in each scenario for the number of ads produced and number of experimental subjects, we simulate 10,000 potential experiments on ads with true effects that vary according to the level of ad variability in the simulation; determine which ads campaigns would run, based on the estimates they would reach of each ads’ effectiveness, which are subject to a degree of sampling error determined by the sample sizes they use in their experiments in that simulation; and finally, calculate how many votes their media campaigns yield given the selected ad’s true effect, assuming that they spend the remainder of their media budget running that ad. These simulations also assume the parameters in Table 3. Finally, for each budget and value of $\frac{\epsilon}{\mu}$, we determine which parameters (number of ads produced and number of experimental subjects) together result in the highest expected vote gain, revealing how much money it would be optimal for campaigns to spend on experiments for any given budget and degree of ad variability.

Figure 4a shows how much money it would be optimal to spend on experiments at various campaign budget sizes and values of $\frac{\epsilon}{\mu}$. First, the line and points at the bottom of the figure show that, at low levels of ad variability ($\frac{\epsilon}{\mu}$), it is optimal for campaigns to spend none of their money on experimentation. However, if ad variability is higher, it is optimal for campaigns to spend a substantial portion of their budgets on experimentation. For instance, under the specific assumptions governing these simulations, if ad variability is the size we estimate in this article, 0.51, campaigns should spend approximately 10%–13% of their media budgets on experimentation. If ad variability were even larger, they should spend even more, approximately 13%–17%. This analysis shows that ad variability is a crucial

TABLE 3. Table of Values Used in Simulations of Ad-Testing Impact in Figure 4. Values Are Based on Assumptions about a Typical Competitive US Senate Campaign

| | Value used in simulations | Source |
|----------------------------|---------------------------|------------------------------------|
| Typical advertising effect | \$200 per net vote | Sides, Vavreck, and Warshaw (2021) |
| Cost of producing one ad | \$15,000 | Practitioners |
| Cost of ad testing | \$2.50 per subject | Practitioners |

FIGURE 4. The Estimated Returns to Experimentation



Note: See Appendix C of the Supplementary Material for further details on estimation.

determinant of how much money campaigns should invest in experimentation.

A corollary of this simulation is that if campaigns under-estimate the degree of ad variability—as we show in Appendix D of the Supplementary Material that many may without our evidence—and therefore under-invest in ad experiments, they will win fewer votes. Appendix C.3 of the Supplementary Material explores this idea in more detail, showing that if campaigns have mistaken beliefs about ad variability, their ad campaigns would likely produce fewer votes. This is

particularly true if they under-estimate ad variability and therefore under-invest in experimentation.

Our second set of simulation results explore the implications of ad testing for campaigns and elections by comparing how many votes campaigns would win if they tested their ads to how many votes they would win if they did not test their ads. These results are based in the same simulations used to construct Figure 4a, except we now estimate how many votes a campaign’s media spending would produce under the scenarios when the campaign (a) spends the optimal amount on

experiments, as determined by the previous simulations, versus (b) if they did not experiment at all. To form the baseline effect of ad spending without experiments, we again follow Sides, Vavreck, and Warshaw's (2021) estimate that ad spending produces votes at \$200 per vote.⁸

Figure 4b shows the results. Unsurprisingly, campaigns with larger budgets (horizontal axis) earn more votes (vertical axis) from spending those larger budgets. However, the use of experimentation does not produce a simple “intercept shift” upward that benefits all campaigns equally. Rather, as shown by the steeper slope of the top line, campaigns with large budgets are better-positioned to reap these rewards: in a world with experimentation, money translates into votes more quickly.

Finally, Figure 4c shows that the extent of ad variability ($\frac{\sigma}{\mu}$) critically conditions both the overall payoffs of experimentation and the differential payoffs for well-financed campaigns. In Figure 4c, the horizontal axis is ad variability. We place a vertical line at our estimated value of 0.51 (with a 95% shaded confidence interval). On the vertical axis, we show that the dollars per vote campaigns would achieve from their total media budget assuming they spend the optimal amount of that budget on experimentation (given the extent of ad variability). The three lines correspond with campaigns of varying overall budgets.

We highlight three main takeaways from Figure 4c. First, part of why experimentation helps wealthier campaigns is the direct effect that, when experimentation is available, ad spending is more effective. Figure 4c shows that, at our estimated level of ad variability of 0.51, all campaigns win votes more cheaply than the \$200/vote benchmark at which our simulations assume ads can win votes in the absence of experimentation (Sides, Vavreck, and Warshaw 2021). Second, experimentation especially benefits wealthier campaigns; when ad variability is 0.51, wealthier campaigns win votes around \$50 per vote less than small campaigns. This difference arises because wealthy campaigns are able to invest more in experimentation. Third, however, the size of the advantage experimentation confers to all campaigns depends on ad variability. When ad variability is zero, all ads are just as persuasive as the average ad, so it does not matter whether campaigns are able to experiment or not. However, in the presence of even low levels of ad variability, campaigns can use experiments to turn money into votes more efficiently—and this decline in cost per vote is especially pronounced for better-financed campaigns.

In summary, scholars have traditionally seen experimentation as a sideshow to the main business of campaigning that can occasionally yield interesting data for

researchers. These findings suggest that experimentation may have much more significant implications for campaigns and for democracy than scholars have previously appreciated. Campaigns now also conduct social science—and campaigns that invest more in experimentation may have a much larger impact on elections. Moreover, elections featuring experiments may be disproportionately won by the candidates and organizations who have the resources to act on the knowledge their experiments provide.

DISCUSSION

In political campaigns, one primary goal of politicians and the political advertising experts they employ is to convince people who otherwise would vote for another candidate to vote for them—they aim to persuade. One of campaigns' principal methods for persuasion is paid television and digital video advertising, but not all ads are likely to be similarly persuasive. In this article, we demonstrated small but meaningful variation in the persuasive effects of advertisements, and argued that this variation has important implications for campaigns and democracy given the advent of advertising experimentation by political groups.

We supported this argument with a unique archive of experiments conducted by campaigns. Importantly, we had access to the entire universe of conducted studies, and these studies were conducted using campaigns' real ads and tested among real voters in the relevant electoral geographies, during live electoral campaigns.

Because of the unprecedented size of this set of experiments, we can describe the distribution of ads' persuasive effects with uncommon precision. We found that the ratio of ads' average effects to the standard deviation of these effects was approximately 0.51, meaning intuitively that it is commonplace for ads to be 51% better than the average ad. Despite the small effects of most ads, our simulations imply serious gains to be had from picking more effective ads when they are shown to millions of people.

However, we also found that it is difficult to predict this variation from observable features of the ads, even when relying on influential theories of advertising effectiveness to guide our expectations. We engaged large teams of research assistants to code the ads on theoretically important dimensions like messenger, message, tone, ask, and production value. Our investigation shows that sometimes these features do seem to be associated with higher effects, but that these associations are inconsistent across contexts. In other words, we were unable to find evidence for general principles about what makes a political ad persuasive in context—what is effective one election year might not be effective the next. There may be no “shortcuts” that allow campaigns to understand which ads will be most persuasive, absent experiments to offer these insights.

These results resonate with findings in other social scientific domains. Milkman et al. (2022) consider the effectiveness of 22 messages to promote flu vaccination and find all positive effects that range from 1 to

⁸ When calculating the implications of running experiments, as noted earlier and considered in further detail later, we also again assume that treatment effects of each ad measured in ad experiments scale down proportionally from the survey environment to the field. We assume true $\frac{\sigma}{\mu} = 0.51$.

3 percentage points. O’Keefe and Hoeken (2021) meta-analyze 30 meta-analyses of message variations from which they conclude “the effect of a given design choice varies considerably from one application to another.” Berman and Van den Bulte (2021) consider 4,964 experiments on website design and find that the average effect size is -0.001 standard units, with a standard deviation of 0.043 units. These three articles have in common with ours that they examine a large pool of treatment effect estimates; they find that effects vary by small but meaningful amounts; and they find that predicting which effects are larger and smaller *ex ante* is difficult.

Our article represents one of the first systematic studies of the payoffs of experimentation for campaigns, and accordingly has several limitations that should prompt future research on this topic. First, although we applied a correction to map the overall magnitude of our survey-based estimates to a field context (see Appendix C.1 of the Supplementary Material), our results depend on the assumption that the dispersion in advertising we observed in a survey context would also be reflected in real-world persuasion. If the survey experimental estimates are biased due to differential attrition, sample nonrepresentativeness, or experimental demand effects, our estimates of the distribution of effects could be incorrect. We have argued why we think these potential sources of bias do not undermine our conclusions, though we underscore that this is an important area for future research.

A second and related assumption of our theoretical analysis is that the effects of ads measured in surveys before an election proportionally “scale down” to their effects on election-day vote choice. Campaigns conduct survey-based ad testing based on this assumption, but this assumption has never been tested to the best of our knowledge, and is very difficult to test because doing so would require estimating the effects of a large number of different ads in the field.⁹ For instance, there may be differences between in-survey estimates and in-field estimates such as the presence of demand effects or the composition of samples that lead effects of different advertisements to not scale down proportionally across modes. Reassuringly, evidence from O’Keefe (2021) suggests that which messages most effectively persuade in surveys is highly predictive of which messages are most effective on behavioral outcomes; Coppock and Green (2015) similarly find strong correspondence between lab and field results. However, these questions too merit further investigation in future studies of survey- and field-study correspondence.

⁹ To see how this assumption might fail, consider the case of decay in effects: some ads might be more effective initially, but others might have effects that last longer. In principle, campaigns might be better off running ads that perform more poorly in in-survey tests with immediately measured outcomes but have greater staying power. Unfortunately, there is fairly little evidence in political science or more generally about heterogeneity in decay across treatments (for one exception, see Coppock 2023, chap. 6), and our research underscores this as an important area for research.

Third, although we tested a number of influential ideas about persuasive approaches, we cannot rule out the possibility that other features of advertisements we did not test might predict which ads persuade better or more consistently.

Fourth, because we necessarily restrict our analysis to ads that were produced, we end up conditioning on a variable (production) that is causally downstream from what we are trying to study (theories of persuasion). Since our goal was to ascertain if ad features could predict ad persuasiveness among the observed set of ads, not estimate the causal effects of those features, this limitation does not invalidate the foregoing analysis. Future investigations of the causal factors that determine ad persuasiveness should follow the lead of Blumenau and Lauderdale (2024) and explicitly randomize each feature separately, though this task is obviously quite daunting.

Fifth, our conclusions about the value of randomized experimentation for choosing more persuasive ads depend critically on an assumption that results from an experiment at least generalize within a given election. Some partial evidence in favor of this assumption comes from Coppock, Hill, and Vavreck (2020), who find similar treatment effects of seven different ads, each measured at multiple points in the election cycle. In Dataverse Appendix DA.3, we also provide an extensive reanalysis of data from several recent articles that tested the effects of policy ads and then retested the ads after 6 months. Our reanalysis shows that ads that performed better in the first test also tended to perform better in the second. These reassuring findings notwithstanding, future work should consider the extent of within-cycle generalizability, as it is a crucial parameter underpinning the logic of experimentation for campaigns.

Sixth, as with all studies, our conclusions may be limited to the time and place where our studies were conducted: with Democratic or left-leaning ads, using video, testing in online-based survey experiments, in the years 2018 and 2020, and in competitive US elections. Indeed, one of our article’s primary contentions is that conclusions about persuasion are likely to differ across contexts. With respect to time, some data suggest that the dynamics of campaign advertising have not changed dramatically: in particular, Sides, Vavreck, and Warshaw (2021, Appendix I) find that TV ad effects have not declined in recent election cycles. Moreover, Green and Platzman (2022) find that rates of partisan change were similar during the Trump administration as during previous eras. Nevertheless, the only way to be sure our study’s findings replicate over time, context, and treatment medium is for future research to conduct such replications, and we welcome such efforts.

Seventh, campaigns may also qualitatively learn from experiments, leading them to change how they produce future advertisements or even to alter their organizational structure to better take advantage of experiments’ insights. Consistent with such a phenomenon, in another context, Koning, Hasan, and Chatterji (2022) find evidence that businesses which experiment

on their products perform better, in part because experiments lead to broader organizational learning. Future research should therefore replicate our research as campaigns' experimentation practices change, and qualitatively investigate how experiments are integrated into and influence campaign practice.

With respect to place, US elections are distinctive in many ways, especially owing to the large amount of money spent in US elections.¹⁰ US electoral districts are also larger than electoral districts in many other countries. Such differences could alter our conclusions with respect to a number of points, such as whether it is optimal for campaigns to spend a meaningful share of their budgets on experimentation, or to what extent ad effects vary between ads. More generally, the fact that Americans are more “Western, Educated, Industrialized, Rich, and Democratic” or “WEIRD” (Henrich, Heine, and Norenzayan 2010) offers many reasons why our results may vary elsewhere. Our article provides a framework for investigating questions related to the impacts and payoffs of experimentation that nevertheless may prove helpful in other contexts.

In addition, as elaborated above, we have incomplete data on differential attrition across studies, although in the subset of studies where we have these data we do not find evidence of asymmetric attrition due to survey drop-out.

Finally, our analysis focused on the implications of experimentation for elections and campaigns, but it is possible that politicians might also learn how to “explain” their positions from randomized experiments (Fenno 1978; Grose, Malhotra, and Van Houweling 2015), suggesting that future research could investigate possible representational consequences of such experimentation.

What are the broader implications of our results for campaigns and for democracy? Seen in one light, these results seem to suggest that randomized experiments have not made good on their purported ability to discover “what works.” Our interpretation is different. The results of the unique archive of randomized experiments we analyze indicate that “what works” to persuade voters changes over time and across contexts. This dynamic suggests that within-election-cycle experimentation may be campaigns' best chance of determining which of their ads are most persuasive. Our analysis shows that the returns to such experimentation for campaigns may be quite substantial. This conclusion has a troubling implication for democracy: the rise of campaign experimentation (Issenberg 2012) suggests that campaigns may increasingly have success at findings ads that persuade, but the benefits of this technology accrue principally to the campaigns with the financial resources to deploy the ads they select at the greatest scale.

¹⁰ For example, *The Hill* recently concluded that “Americans spend more on politics and political campaigns than any other nation on Earth. ...U.S. spending is such that it's barely comparable to the amounts in other countries,” <https://thehill.com/homenews/campaign/529080-us-election-spending-exceeds-gdp-of-numerous-countries/>.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <https://doi.org/10.1017/S0003055423001387>.

DATA AVAILABILITY STATEMENT

Replication data are available at the American Political Science Review Dataverse: <https://doi.org/10.7910/DVN/LBPSSV>.

ACKNOWLEDGMENTS

We are grateful to Joshua Kalla for many contributions to this project, including analysis of the 2018 data. We thank Aliesa Bahri, Nick Begotka, Avery Brown, Jesse Bryant, Claire Calkins, Jamie Druckman, Stella Fitzgerald, Don Green, Spencer Hagaman, Kevin Han, Shanto Iyengar, Ananya Kachru, Ben Lauderdale, Carrie Mannino, Greg Martin, Jon Mellon, Dan O'Keefe, Isabel Rooper, Zachary Peskowitz, Olivia Probst, Mitchell Mares, Hovik Minasyan, Lucio Moscarini, Kevin Munger, TJ Noel-Sullivan, Emma Mueller, John Sides, Ellie Singer, Aprajita Singh, Lynn Vavreck, Chris Warshaw, and Robb Willer for their excellent comments and feedback. We thank Phoenix Dalto for performing an independent verification of our experimental analysis and Mike Baumer, Kathy Gerlach, Tanya Martin, Kwaku Ofori-Atta, and Josh Dean at Swayable for their assistance. We are grateful to participants at an EGAP feedback session for their expert guidance. We thank audiences at the London School of Economics and Political Science and the MEAD workshop at the University of Wisconsin–Madison for their comments and criticisms.

AUTHOR CONTRIBUTIONS

Swayable (J.S., V.C., N.L., and M.H.) designed and conducted all experiments collected in this article, then provided raw data to L.H., D.B., A.C., and B.M.T. for independent analysis, with express permission to publish any findings. This study is governed by two pre-analysis plans. We filed the first with the code we planned to use to analyze the data from the 2018 experiments: https://osf.io/q276a/?view_only=0a7ee62f8ede464ea416e4cc9e3c2637. After observing the results of the 2018 analysis but before analyzing the data from the 2020 experiments, we then filed a follow-up pre-analysis plan for the 2020 data: https://osf.io/5c9hx/?view_only=f728d3cbc8b848dfa6bc02402828750c. The two PAPs are largely similar; we follow the 2020 PAP when analyzing the data, creating small deviations from that 2018 PAP that we note where applicable. Some methods used in this study are based on US patent US011756061B2, developed and owned by Swayable.

CONFLICT OF INTEREST

J.S., V.C., N.L., and M.H. declare that they have a financial interest in Swayable. L.H. and B.M.T. are founders of a research organization that conducts public opinion research.

ETHICAL STANDARDS

The authors affirm that this research did not involve human subjects.

REFERENCES

- Albertson, Bethany, Lindsay Dun, and Shana Kushner Gadarian. 2020. "The Emotional Aspects of Political Persuasion." In *The Oxford Handbook of Electoral Persuasion*, eds. Elizabeth Suhay, Bernard Grofman, and Alexander H. Trechsel, 169–83. New York: Oxford University Press.
- Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88 (4): 829–38.
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.
- Berman, Ron, and Christophe Van den Bulte. 2021. "False Discovery in A/B Testing." *Management Science* 68 (9): 6355–7064.
- Bigsby, Elisabeth, and Samuel R. Wilson. 2020. "Reactance." In *The International Encyclopedia of Media Psychology*, 1–5. Hoboken, NJ: Wiley-Blackwell.
- Blumenau, Jack, and Benjamin E. Lauderdale. 2024. "The Variable Persuasiveness of Political Rhetoric." *American Journal of Political Science* 68 (1): 255–70.
- Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. Hoboken, NJ: John Wiley & Sons.
- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49 (2): 388–405.
- Broockman, David E., and Joshua L. Kalla. 2022. "When and Why Are Campaigns' Persuasive Effects Small? Evidence from the 2020 US Presidential Election." *American Journal of Political Science* 67 (4): 833–49.
- Coppock, Alexander. 2023. *Persuasion in Parallel*. Chicago, IL: University of Chicago Press.
- Coppock, Alexander, and Donald P. Green. 2015. "Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3 (1): 113–31.
- Coppock, Alexander, Seth J. Hill, and Lynn Vavreck. 2020. "The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-Time Randomized Experiments." *Science Advances* 6 (36): eabc4046. <https://doi.org/10.1126/sciadv.abc4046>.
- Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. Indianapolis, IN: Pearson College Division.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5.
- Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105 (1): 135–50.
- Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.
- Green, Donald P., and Alan S. Gerber. 2019. *Get Out the Vote: How to Increase Voter Turnout*, 3rd edition. Washington, DC: Brookings Institution.
- Green, Donald P., Mary C. McGrath, and Peter M. Aronow. 2013. "Field Experiments and the Study of Voter Turnout." *Journal of Elections, Public Opinion and Parties* 23 (1): 27–48.
- Green, Donald P., and Paul Platzman. 2022. "Partisan Stability during Turbulent Times: Evidence from Three American Panel Surveys." *Political Behavior*: 1–27. <https://doi.org/10.1007/s11109-022-09825-y>
- Grose, Christian R., Neil Malhotra, and Robert Parks Van Houweling. 2015. "Explaining Explanations: How Legislators Explain Their Policy Positions and How Citizens React." *American Journal of Political Science* 59 (3): 724–43.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "Most People Are Not WEIRD." *Nature* 466 (7302): 29.
- Hewitt, Luke, David Broockman, Alexander Coppock, Ben M. Tappin, James Slezak, Nathaniel Lubin, et al. 2024. "Replication Data for: How Experiments Help Campaigns Persuade Voters: Evidence from a Large Archive of Campaigns' Own Experiments." Harvard Dataverse. Dataset. <https://doi.org/10.7910/DVN/LBPSSV>.
- Hovland, Carl Iver, Irving Lester Janis, and Harold H. Kelley. 1953. *Communication and Persuasion*. New Haven, CT: Yale University Press.
- Issenberg, Sasha. 2012. *The Victory Lab: The Secret Science of Winning Campaigns*. New York: Crown.
- Iyengar, Shanto, and Nicholas A. Valentino. 2000. "Who Says What? Source Credibility as a Mediator of Campaign Advertising." In *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*, eds. Arthur Lupia, Mathew D. McCubbins, and Samuel L. Popkin, 108–29. Cambridge: Cambridge University Press.
- Jacobson, Gary C., and Jamie L. Carson. 2019. *The Politics of Congressional Elections*. Washington, DC: Rowman & Littlefield.
- Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–66.
- Kendall, Chad, Tommaso Nannicini, and Francesco Trebbi. 2015. "How Do Voters Respond to Information? Evidence from a Randomized Campaign." *American Economic Review* 105 (1): 322–53.
- Koning, Rembrand, Shariqee Hasan, and Aaron Chatterji. 2022. "Experimentation and Start-Up Performance: Evidence from A/B Testing." *Management Science* 68 (9): 6434–53.
- Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. "The Effects of Negative Political Advertisements: A Meta-Analytic Assessment." *American Political Science Review* 93 (4): 851–75.
- Lau, Richard R., Lee Sigelman, and Ivy Brown Rovner. 2007. "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics* 69 (4): 1176–209.
- Martin, Gregory J, and Zachary Peskowitz. 2015. "Parties and Electoral Performance in the Market for Political Consultants." *Legislative Studies Quarterly* 40 (3): 441–70.
- Milkman, Katherine L., Linnea Gandhi, Mitesh S. Patel, Heather N. Graci, Dena M. Gromet, Hung Ho, Joseph S. Kay, et al. 2022. "A 680,000-Person Megastudy of Nudges to Encourage Vaccination in Pharmacies." *Proceedings of the National Academy of Sciences* 119 (6): e2115126119c. <https://doi.org/10.1073/pnas.211512611>.
- Munger, Kevin. 2019. "The Limited Value of Non-Replicable Field Experiments in Contexts with Low Temporal Validity." *Social Media + Society* 5 (3): 2056305119859294. <https://doi.org/10.1177/2056305119859294>.
- Nelson, Phillip. 1974. "Advertising as Information." *Journal of Political Economy* 82 (4): 729–54.
- O'Keefe, Daniel J. 1997. "Standpoint Explicitness and Persuasive Effect: A Meta-Analytic Review of the Effects of Varying Conclusion Articulation in Persuasive Messages." *Argumentation and Advocacy* 34 (1): 1–12.
- O'Keefe, Daniel J. 2021. "Persuasive Message Pretesting Using Non-Behavioral Outcomes: Differences in Attitudinal and Intention Effects as Diagnostic of Differences in Behavioral Effects." *Journal of Communication* 71 (4): 623–45.
- O'Keefe, Daniel J., and Hans Hoeken. 2021. "Message Design Choices Don't Make Much Difference to Persuasiveness and Can't

- Be Counted On—Not Even When Moderating Conditions Are Specified.” *Frontiers in Psychology* 12: 664160. <https://doi.org/10.3389/fpsyg.2021.664160>
- O’Keefe, Daniel James. 2004. “Trends and Prospects in Persuasion Theory and Research.” In *Readings in Persuasion, Social Influence, and Compliance Gaining*, 31–43. Boston, MA: Pearson/Allyn and Bacon.
- Schlesinger, Mark, and Richard R. Lau. 2000. “The Meaning and Measure of Policy Metaphors.” *American Political Science Review* 94 (3): 611–26.
- Searles, Kathleen, Erika Franklin Fowler, Travis N. Ridout, Patricia Strach, and Katherine Zuber. 2020. “The Effects of Men’s and Women’s Voices in Political Advertising.” *Journal of Political Marketing* 19 (3): 301–29.
- Sides, John, Lynn Vavreck, and Christopher Warshaw. 2021. “The Effect of Television Advertising in United States Elections.” *American Political Science Review* 116 (2): 702–18.
- Strach, Patricia, Katherine Zuber, Erika Franklin Fowler, Travis N. Ridout, and Kathleen Searles. 2015. “In a Different Voice? Explaining the Use of Men and Women as Voice-Over Announcers in Political Advertising.” *Political Communication* 32 (2): 183–205.
- Thibodeau, Paul H., and Lera Boroditsky. 2011. “Metaphors We Think With: The Role of Metaphor in Reasoning.” *PLoS One* 6 (2): e16782. <https://doi.org/10.1371/journal.pone.0016782>
- Thurber, James A., and Candice J. Nelson. 2001. *Campaign Warriors: Political Consultants in Elections*. Washington, DC: Brookings Institution.
- Vavreck, Lynn. 2001. “The Reasoning Voter Meets the Strategic Candidate: Signals and Specificity in Campaign Advertising, 1998.” *American Politics Research* 29 (5): 507–29.
- Vavreck, Lynn. 2009. *The Message Matters*. Princeton, NJ: Princeton University Press.
- Weber, Christopher, Johanna Dunaway, and Tyler Johnson. 2012. “It’s All in the Name: Source Cue Ambiguity and the Persuasive Appeal of Campaign Ads.” *Political Behavior* 34 (3): 561–84.
- Yourman, Julius. 1939. “Propaganda Techniques Within Nazi Germany.” *Journal of Educational Sociology* 13 (3): 148–63.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.