Online Appendix

Please note that per the Editors' instructions to us regarding the APSR Online Appendix page limit guidelines, some content in the Online Appendix was moved to a supplementary Dataverse Online Appendix. Appendix sections, figures, and tables beginning with DA (e.g., section DA1, Table DA4, etc.) can be found in the Dataverse Online Appendix on Dataverse.

A Appendix: preregistered analyses

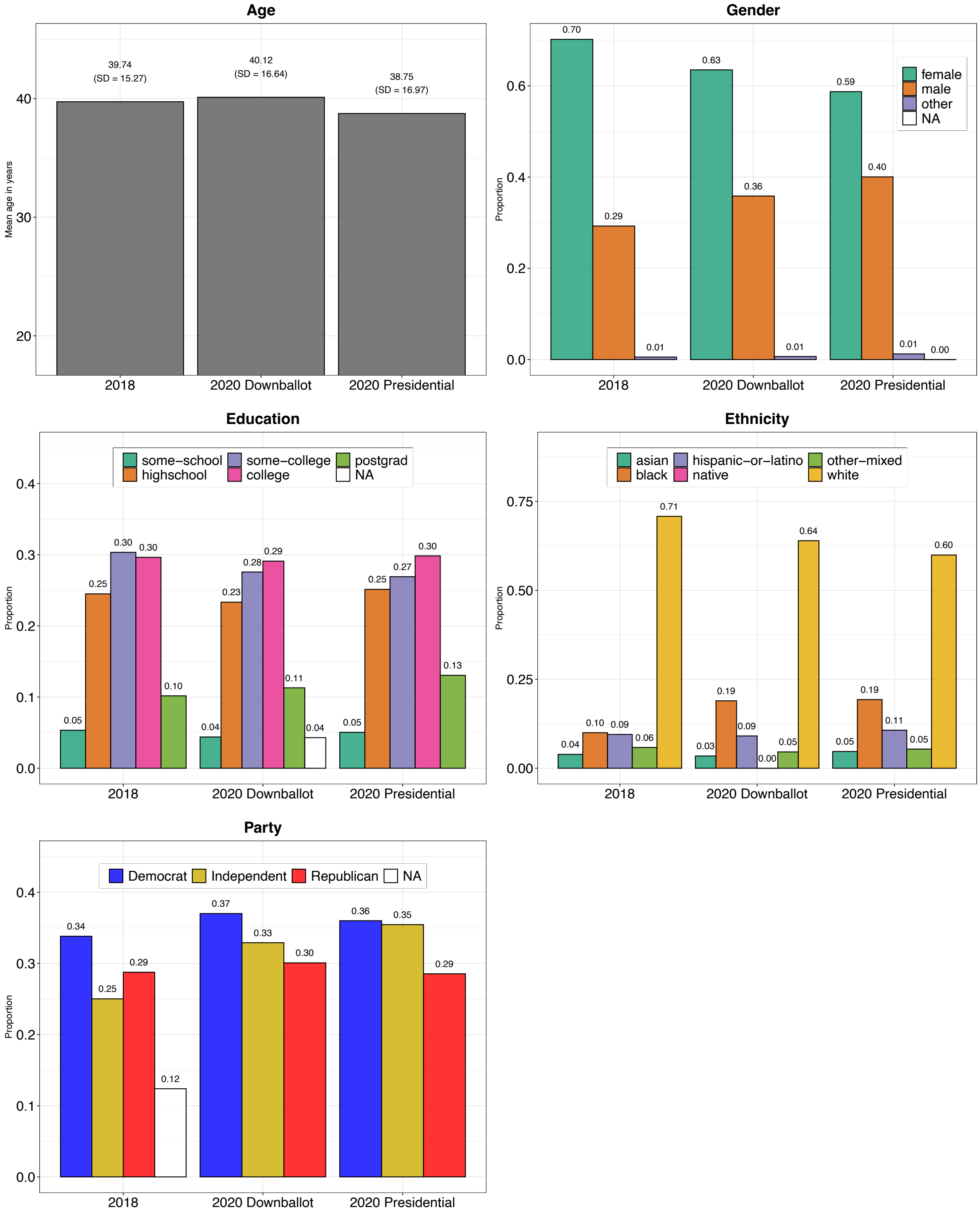
A.1 Demographics

The pretreatment covariates were measured with these questions:

- What is your year of birth?
- What is your gender *(dummy coded: Male/Female/Other)*
- What race or ethnic group do you most identify with? *(dummy coded: White, Black, Asian, Hispanic/Latino, Other/mixed)*
- What is your educational background? *(dummy coded: Some School/No Diploma, High School Graduate, Some College, College Degree, Postgraduate Degree)*
- In terms of politics, do you consider yourself a Democrat, independent, or Republican? *(0-10)*
- On a scale from very liberal to very conservative, how would you best describe your political views? *(0-10)*
- Do you approve or disapprove of the way Donald Trump is handling his job as President? *(0-10)*

A summary of the major demographic covariates is described in Figure OA1.

Figure OA1: Respondent demographics.



A.2 Metaregression coefficient plots

In this section, we provide coefficient plots that correspond to the t-statistics reported in Figure 3.

Figure OA2 shows that, in 2020, ads that had policy issues or the candidates themselves as the primary focus did better than ads that did not on the favorability outcome. However, this pattern was not replicated in the 2020 data. Mentions of candidate facts were associated with stronger effects in 2018 and in the 2020 downballot races, but not in the Presidential race. None of the rhetorical techniques we thought might be important – positive transfer of association, positive testimonial, etc – seem to generate larger or smaller persuasive effects. Figure OA3 shows that with few exceptions, the same patterns hold for the vote choice outcome.

Figures OA4 and OA5 show the results for the secondary outcomes. Production value, the messenger being a politician or not, the ad mentioning a new fact, containing an explicit ask for a vote, deploying the emotion of anger, or using a particular town—-none of these are consistently associated with higher treatment effects.

Finally, Figures OA6 and OA7 report the estimates from new hypotheses we introduced in the 2020 PAP. Here again we see a similar story: small, inconsistent differences depending on the messenger of the ad or the issue mentioned.

Figure OA2: **Primary** metaregressions for **favorability** outcome

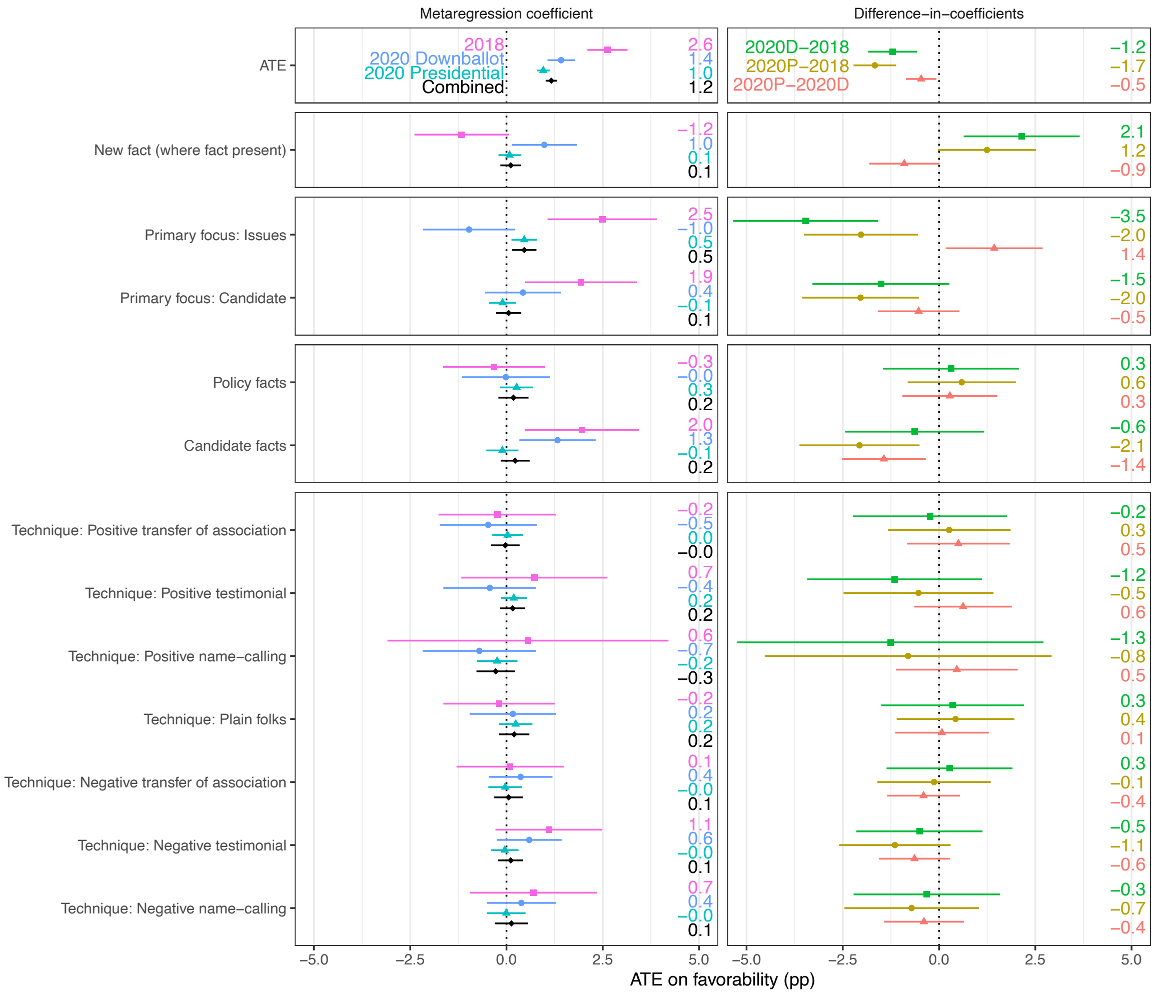


Figure OA3: Primary metaregressions for **votechoice** outcome

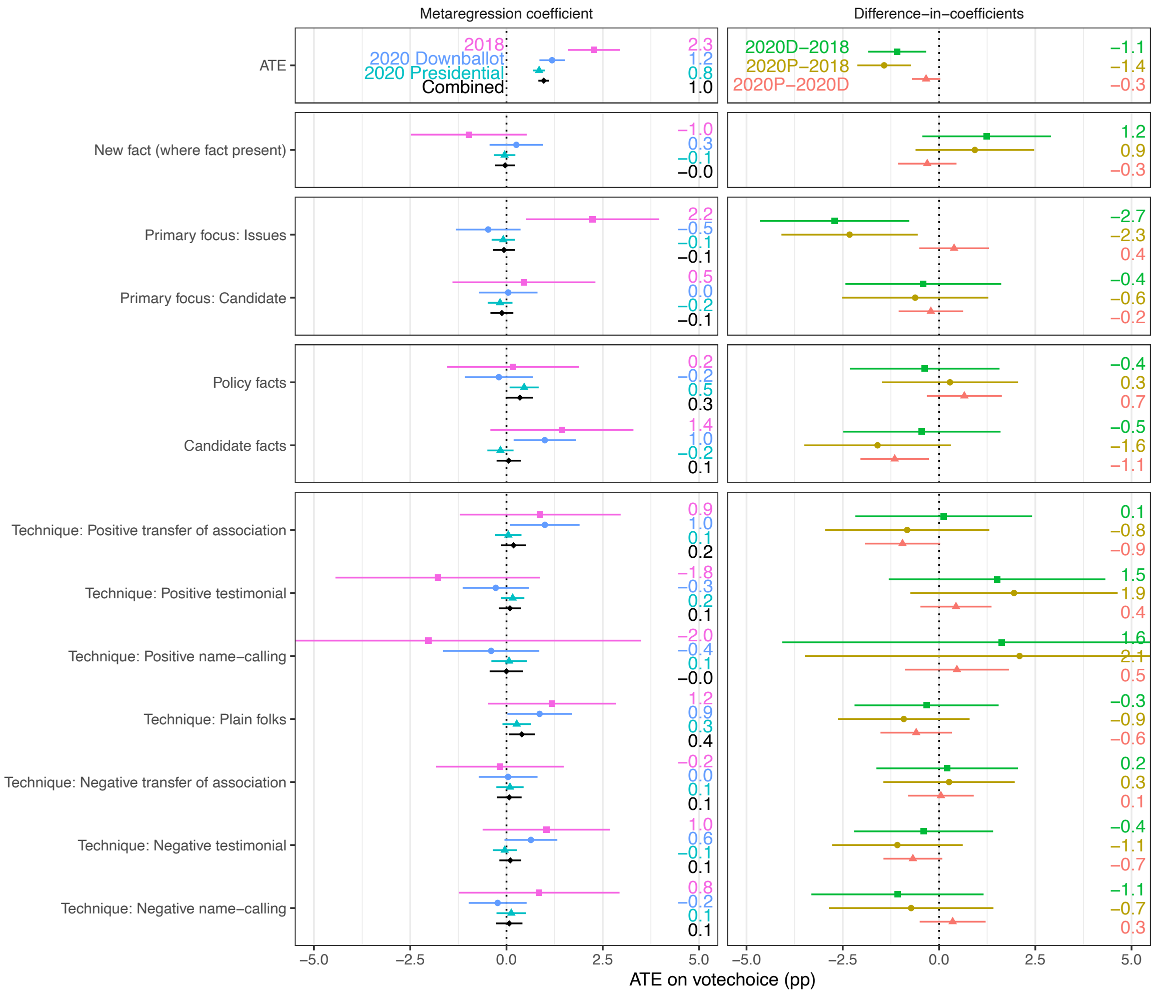


Figure OA4: Secondary metaregressions for **favorability** outcome

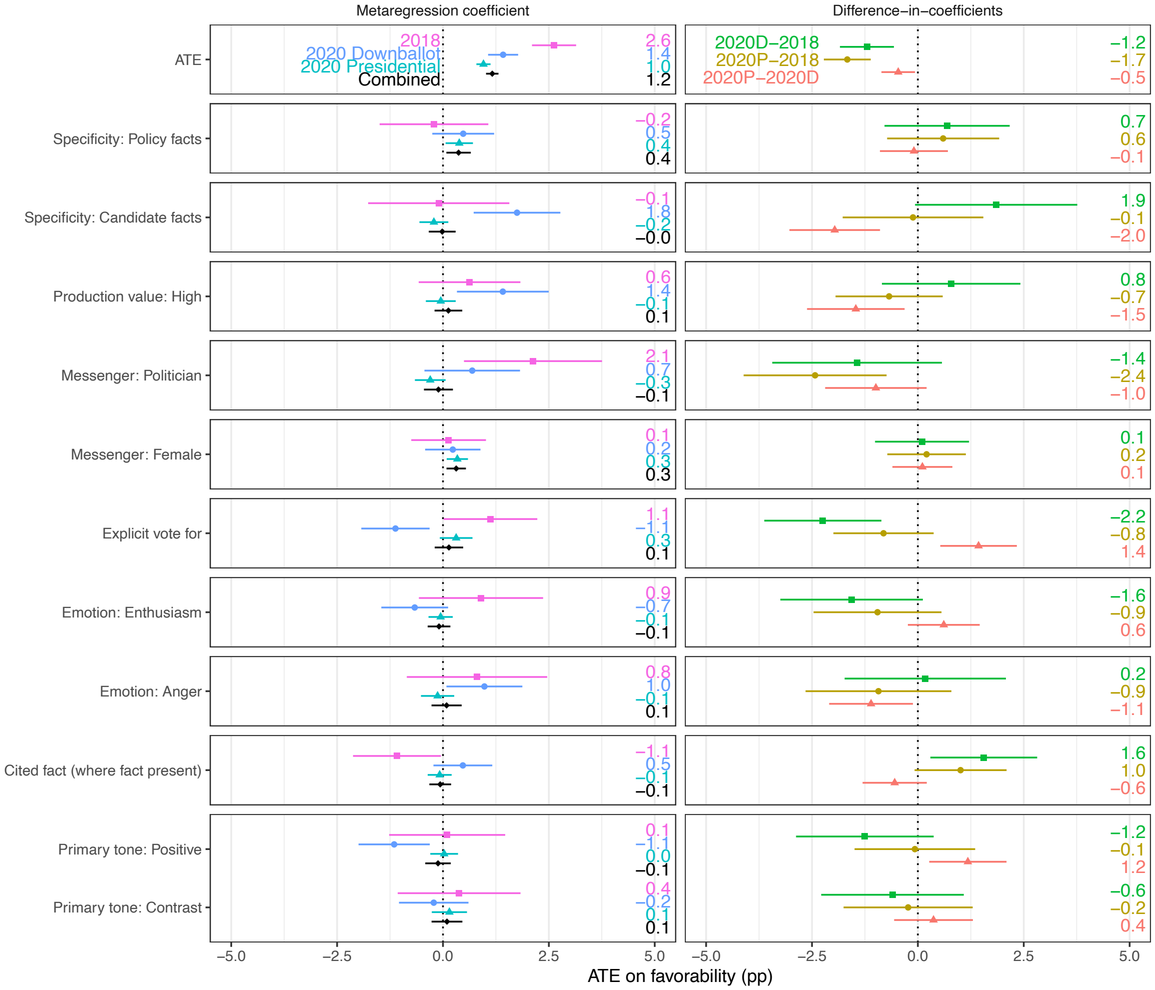


Figure OA5: **Secondary** metaregressions for **votechoice** outcome

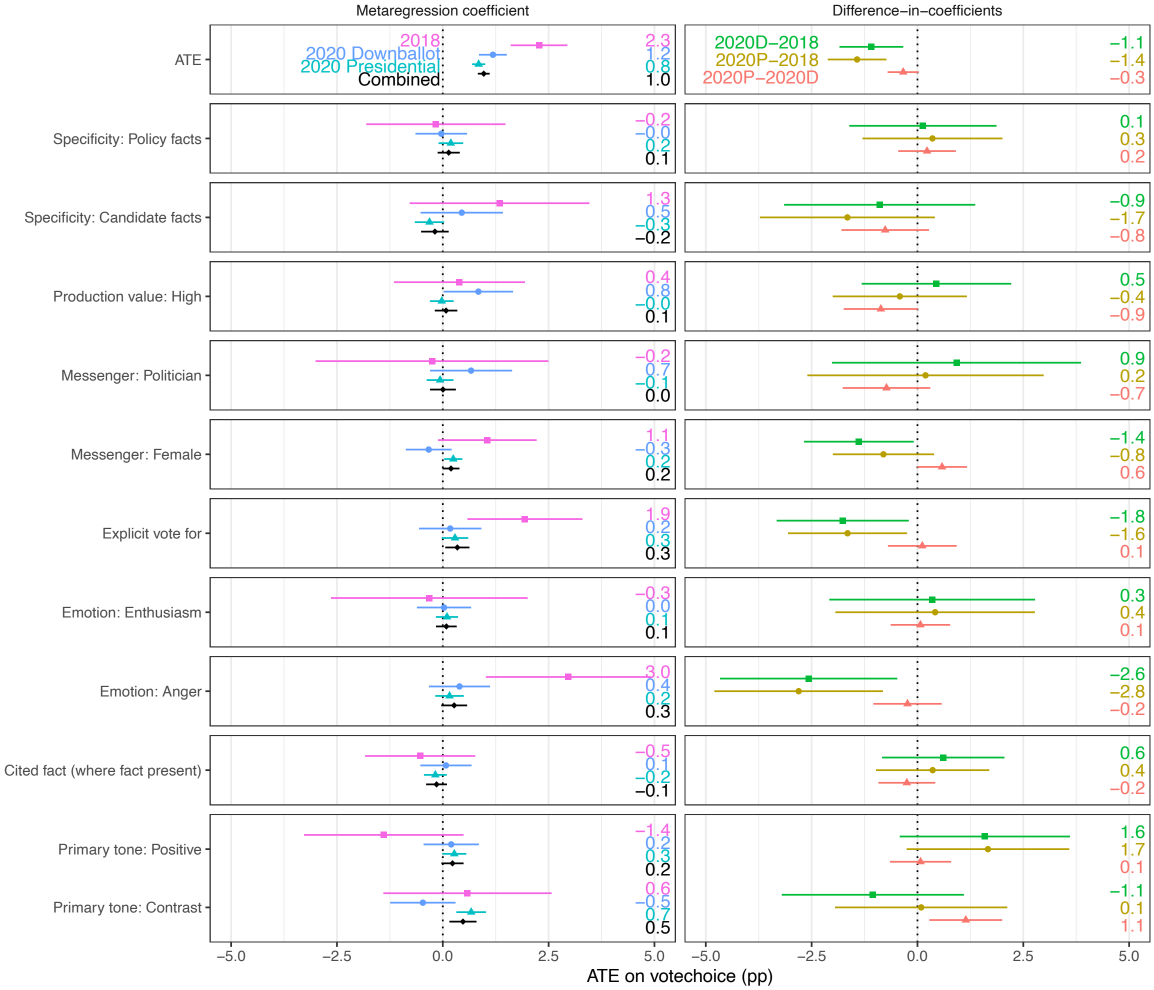


Figure OA6: New metaregressions for **favorability** outcome

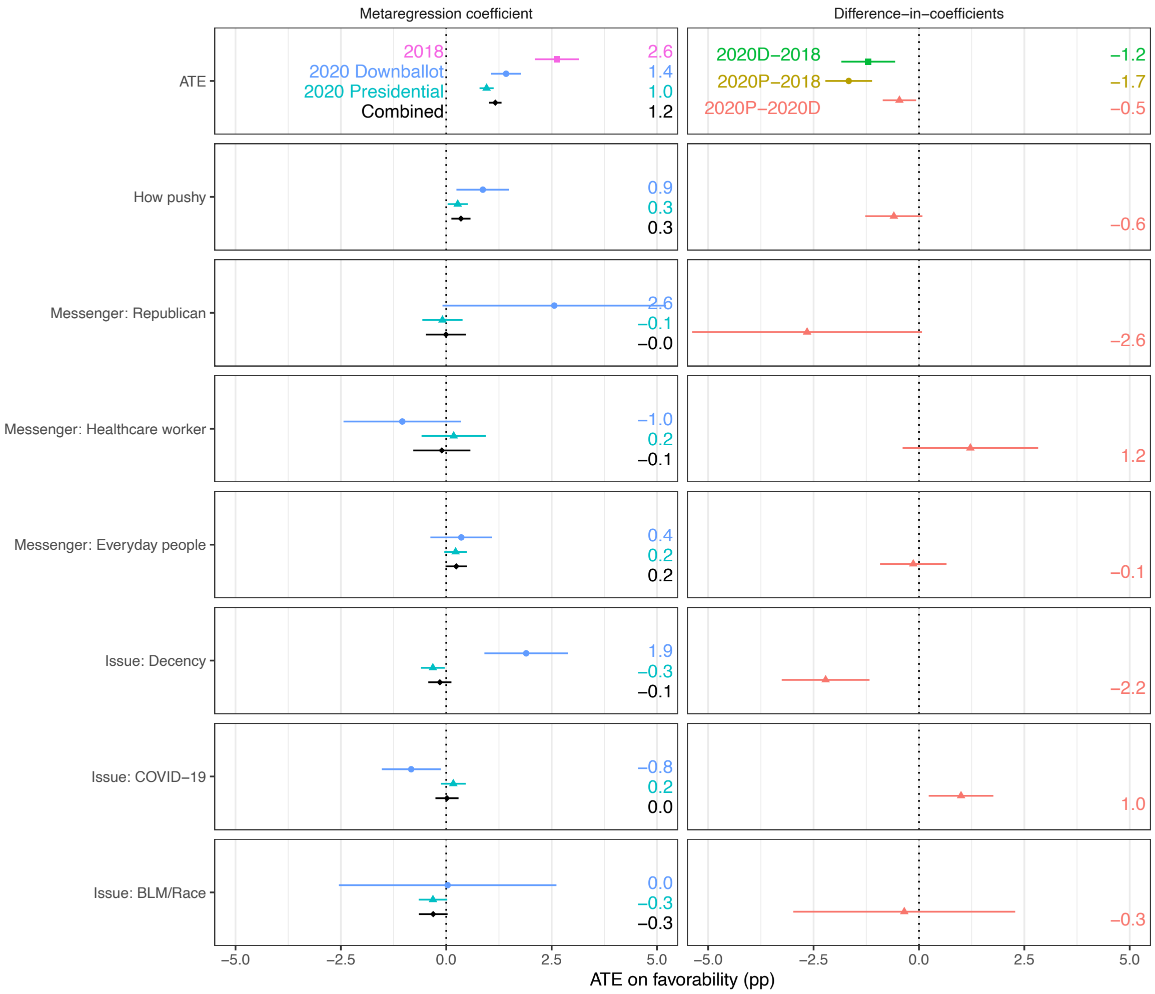
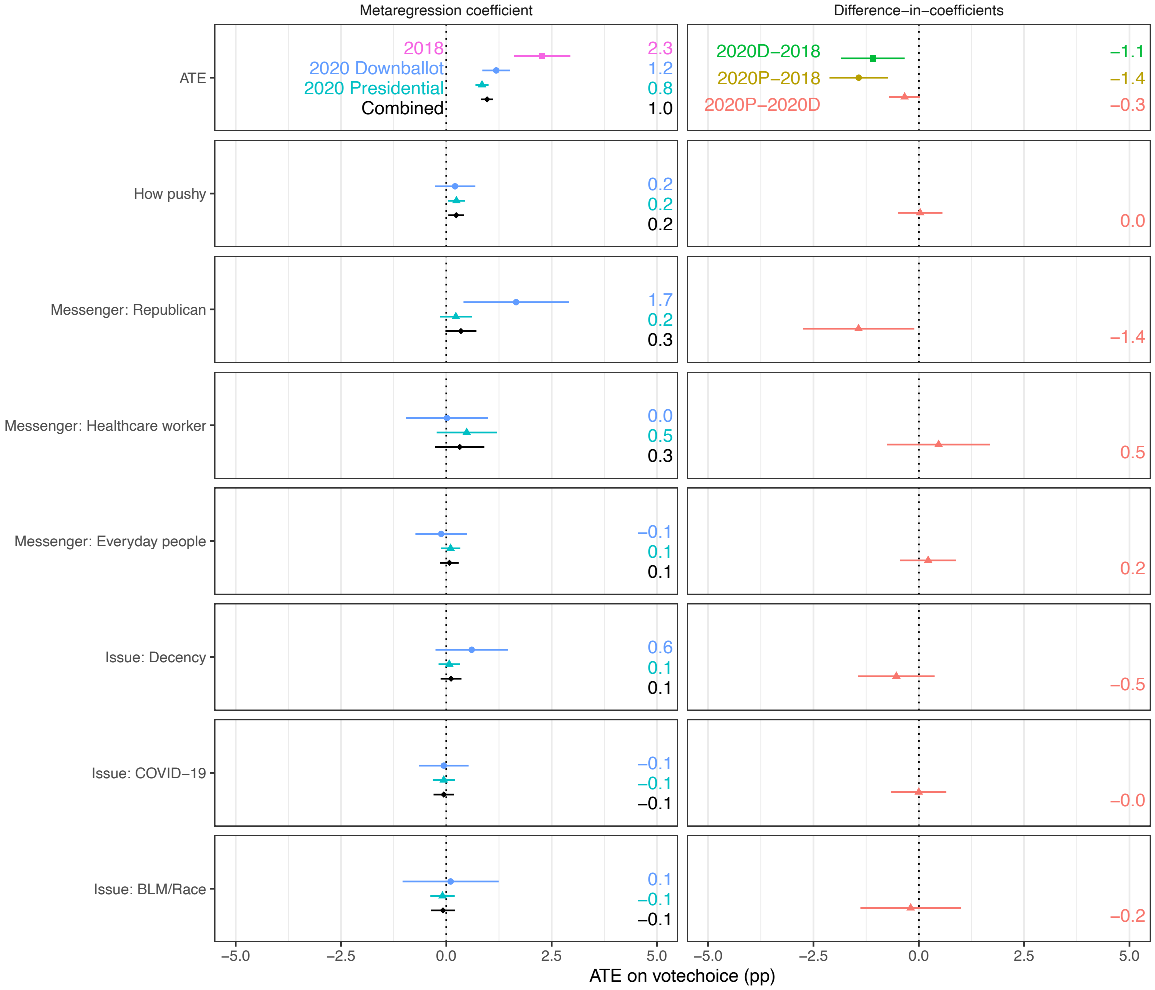
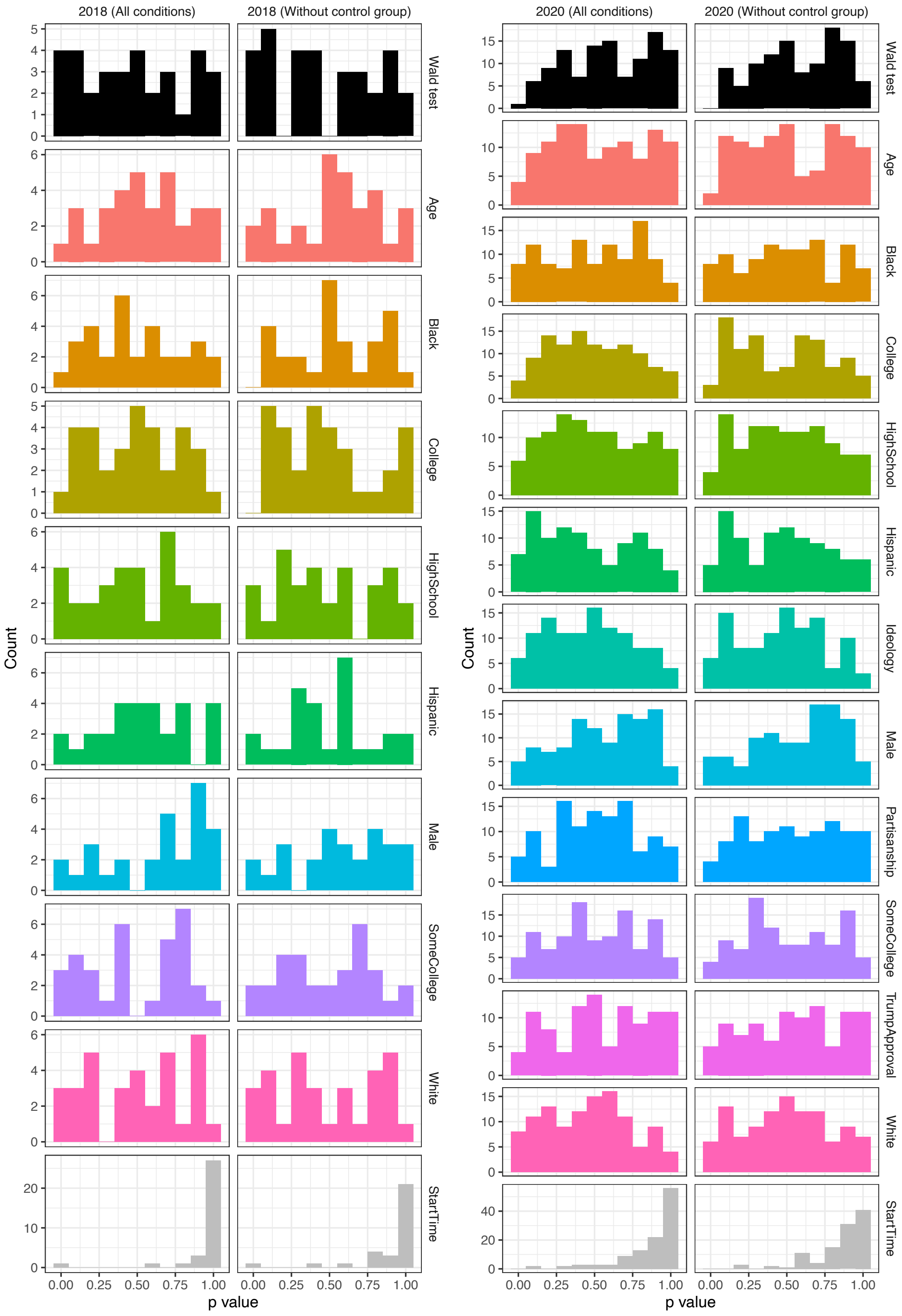


Figure OA7: New metaregressions for **votechoice** outcome



A.3 Balance checks

Figure OA8: Balance checks for all studies analysed. Variables are measured post-treatment in 2018 data, and pre-treatment in 2020 data.



A.4 Reliability

Figure OA9: Reliability of ratings for all video features used in the analysis. This plot shows the estimated consistency of the $k-$rater average measure for each item, $ICC(C, k)$, using $k = 3$ ratings per video in 2018 and $k = 2$ ratings per video in 2020.

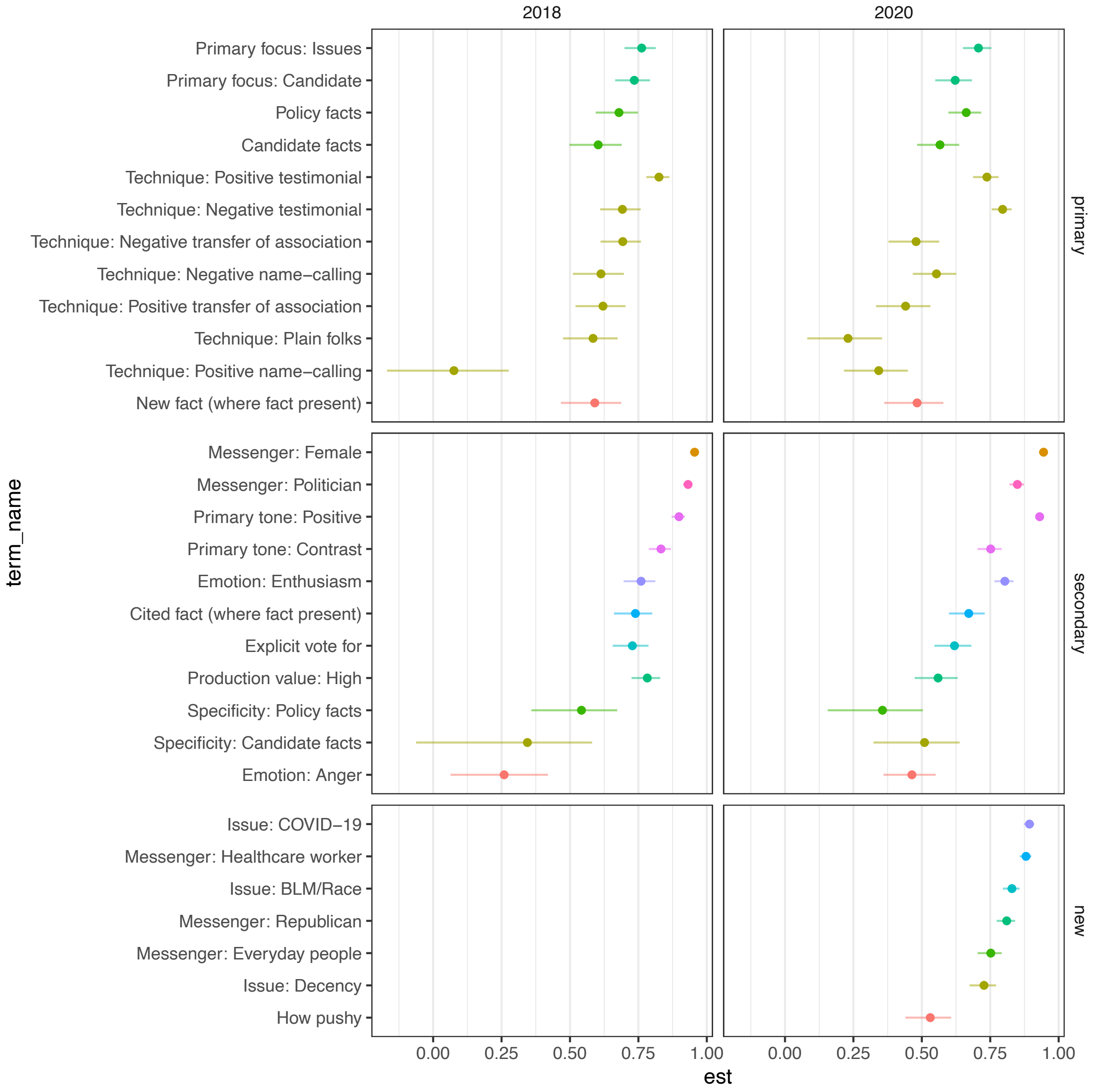
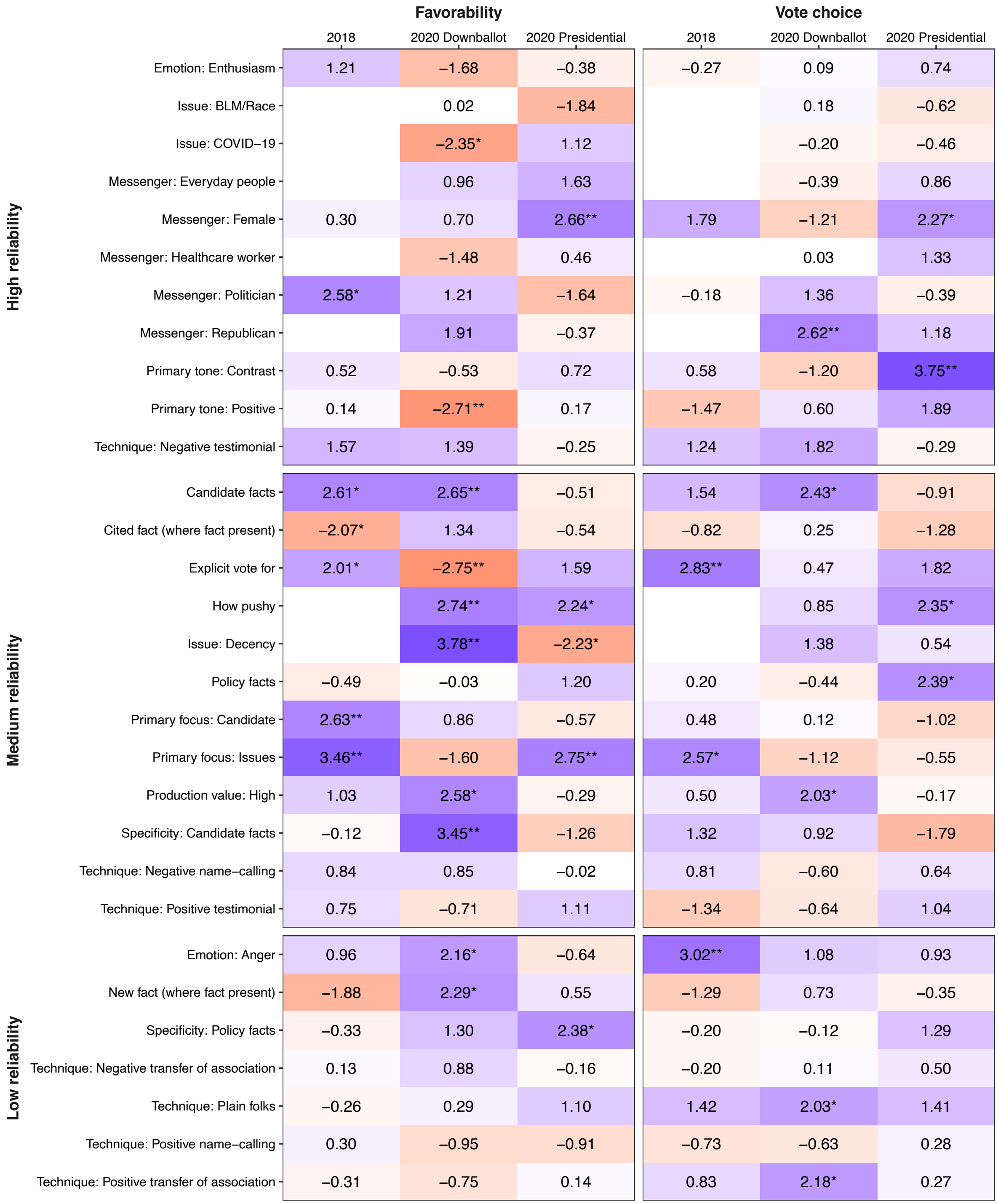


Figure OA10: Metaregression $t-$statistics matrix, arranged by reliability ($ICC(C,k)$) among 2020 raters.



Notes: Notes: Low: $ICC < 0.5$. Medium: $0.5 < ICC < 0.75$. High: $ICC > 0.75$. 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).

B Appendix: Non-preregistered analyses

B.1 Metaregression tables

Table OA1: Metaregressions table "Time to election". See DA 2 for details.

Metaregressions with "Favorability" outcome

Predictor	**2018**	**2020D**	**2020P**	**Combined**	**2020D-2018**	**2020P-2018**	**2020P-2020D**
Days until election (log scale)	**1.8 (0.6)**	**2.0 (0.5)**	0.2 (0.2)	**0.4 (0.1)**	0.2 (0.8)	**-1.5 (0.6)**	**-1.8 (0.5)**
Intercept	**3.0 (0.4)**	**1.7 (0.2)**	**0.9 (0.1)**				
Race: Other		**-2.4 (0.6)**					
Race: Gov	**1.7 (0.7)**						
Race: StateLeg	0.0 (1.0)	**-1.0 (0.4)**					
$\hat{R}^2$ (all vs. control)	0.04	0.13	0.02				$p = 0.000$
$\hat{R}^2$ (all predictors)	0.30	0.16	0.02				
$\hat{\sigma}$	1.40	0.78	0.44				
$N_{treatments}$	131	131	170				

Metaregressions with "Vote choice" outcome

Predictor	**2018**	**2020D**	**2020P**	**Combined**	**2020D-2018**	**2020P-2018**	**2020P-2020D**
Days until election (log scale)	-0.2 (0.6)	0.1 (0.5)	0.1 (0.1)	0.1 (0.1)	0.3 (0.8)	0.3 (0.7)	0.0 (0.5)
Intercept	**2.2 (0.4)**	**1.7 (0.3)**	**0.8 (0.1)**				
Race: Other		-0.5 (0.6)					
Race: Gov	-0.4 (1.2)						
Race: GA Runoff		-0.7 (0.4)					
Race: StateLeg	-0.1 (2.2)	**-1.7 (0.5)**					
$\hat{R}^2$ (all vs. control)	< 0	< 0	< 0				$p = 0.914$
$\hat{R}^2$ (all predictors)	< 0	0.18	< 0				
$\hat{\sigma}$	1.44	0.43	0.35				
$N_{treatments}$	101	181	292				

B.2 Robustness: dichotomized votechoice

Table OA2: Estimated mean and variability in ATEs when using dichotomized vote choice.

	No moderators			Race fixed effects		Study fixed effects	
Election	μ	τ	pval	τ	pval	τ	pval
2018	2.64 $[1.83, 3.45]$	1.67 $[1.06, 2.36]$	$< .001$	1.69 $[1.07, 2.39]$	$< .001$	1.50 $[0.90, 2.17]$	$< .001$
2020 Downballot	1.51 $[0.98, 2.03]$	0.66 $[0.00, 1.29]$	0.069	0.62 $[0.00, 1.26]$	0.136	0.68 $[0.00, 1.33]$	0.074
2020 Presidential	1.07 $[0.81, 1.32]$	0.50 $[0.03, 0.77]$	0.007	0.50 $[0.03, 0.77]$	0.007	0.48 $[0.00, 0.75]$	0.028

B.3 Differential attrition analysis

In this subsection we discuss and present robustness checks regarding differential attrition. To summarize what follows:

- There are two sources of attrition in the data, and the sources are different by year. In the 2018 data, questions were optional, and so respondents could attrit by leaving questions blank *or* leaving the survey. In the 2020 data, questions were mandatory, so respondents could only attrit by leaving the survey.

- In the 2018 data, we are able to measure whether individuals left a question blank, but not whether they left the survey. We see evidence that participants in the control group were more likely to leave the questions blank. However, our analyses which use study fixed effects only examine variation between treatment arms (not the control), and we see limited differential attrition of this form between treatment arms (see second row of Figure OA13).

- In the 2020 data, the questions were mandatory but Swayable originally collected no data on whether the survey was completed. In response to our inquiries they were able to reconstruct data on survey completion for a subset of studies from their raw logs, which began collecting data on incomplete surveys only towards the end of 2020. We see limited signs of differential attrition in this subset of the data in all but one study.

To provide further detail, in addition to overall balance checks presented in Appendix A.3, we test separately for two sources of attrition in our data. First, in the 2018 dataset only, respondents were able to opt out of answering outcome questions. We observe whether this occurred in the 2018 data. Second, in both years, some respondents exit the entire survey experiment post-treatment. For this second kind of attrition, we are only able to analyse a subset of the affected studies (those conducted towards the end of 2020) because, in earlier data, respondents who exited the survey mid-way were not recorded.

In the top row of Figure OA11, we test for attrition-induced covariate imbalance by either mechanism, by regressing $missing{*}covariate$ on treatment condition (where $missing$ is a dummy variable indicating when a respondent did not provide outcome data, and $covariate$ is a demeaned demographic variable). This regression estimates the extent to which attrition produces imbalance on each covariate between treatment arms, similar to the preregistered balance checks presented in Appendix A.3.

These tests reveal that differential non-response in the 2018 dataset produced statistically detectable covariate imbalance between treatment and control groups, but not between different treatments. This is primarily driven by a difference in the overall attrition *rate*, which was typically 5-10% larger in the control group than the treatment group (Figure OA13, top). We hypothesize that this is due to unfamiliarity with the candidates (races in the 2018 dataset were typically low-salience) causing some respondents to opt out when in the control-group. Critically, this means that our specifications with study fixed effects should not be affected by differential attrition.

In addition, while statistically significant, the scale of this covariate imbalance is substantively small and unlikely to materially affect the estimated ATEs in each individual survey, or the subsequent metaregressions in our main analyses without study fixed effects. In Figure OA11 (bottom), we estimate ATEs for each treatment using an IPW regression, in which respondents are inversely weighted by their propensity to provide a response[11]. These estimates differ only minimally from the unweighted estimates used for our preregistered analysis.

Furthermore, in Figure OA12 we present robustness checks in which all metaregressions are re-calculated using these IPW-estimated ATEs, as well as two alternative specifications that include fixed- or random- effects for study id (Note: these latter two specifications are somewhat extreme adjustments, as much of the variability in ad features is itself explained by the study an ad was tested in). We find that all alternative metaregressions specifications preserve the same broad pattern of results as our preregistered specifications, with small deviations relative to the large differences found between datasets. This provides strong evidence that attrition between treatment and control groups in the 2018 data does not substantively alter our main results presented in Figure 3.

Finally, in Figure OA13, we present several other common tests for differential attrition. The columns in this Figure are the year of the study (2018 on the left, where we can only measure missingness due to the questions being optional; and 2020 on the right, for the subset of studies where we have access to missingness due to leaving the survey) and the rows are the following statistical tests:

1. Top row: **Difference in attrition rates between treatment and control groups.** As discussed above, this is substantial in the 2018 data due to opt-out outcomes. However, any attrition in the control group does not affect our analyses with study fixed effects, which examine variation among treatment arms only.

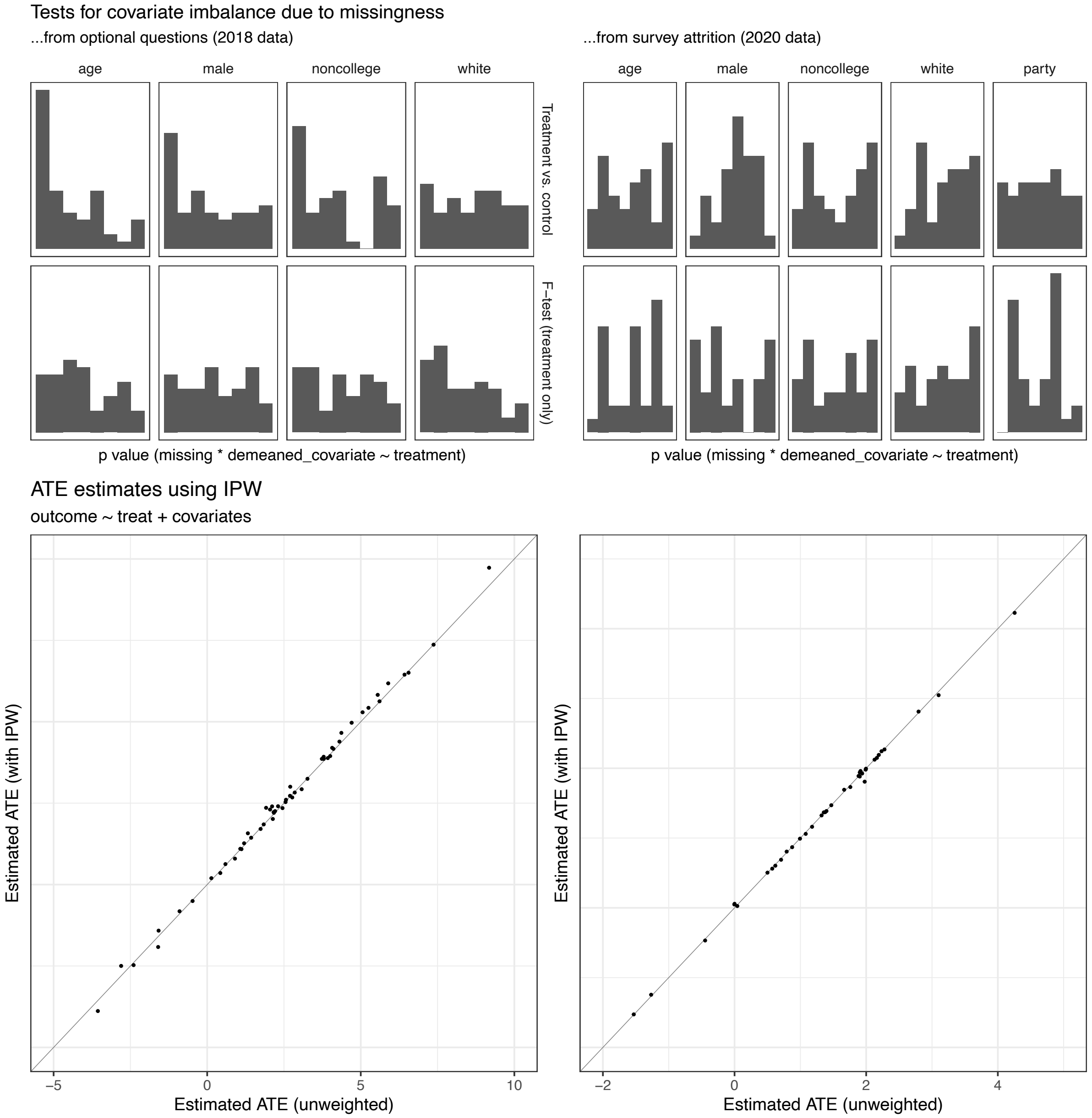
2. Middle row: **F-test on treatment differences.** Restricted to treatment videos only, we test whether the rate of attrition varies by the specific treatment. The p values of these tests (one per study) appear fairly uniform, suggesting that the attrition rate is the same across treatment groups.

3. Bottom row: **F-test on interaction with covariates.** We test for an interaction term between treatment condition and covariates, on the rate of attrition. These tests include the control

[11] This propensity itself is estimated using a logistic regression, $attrit \sim treat * (age + gender + ethnicity + education)$

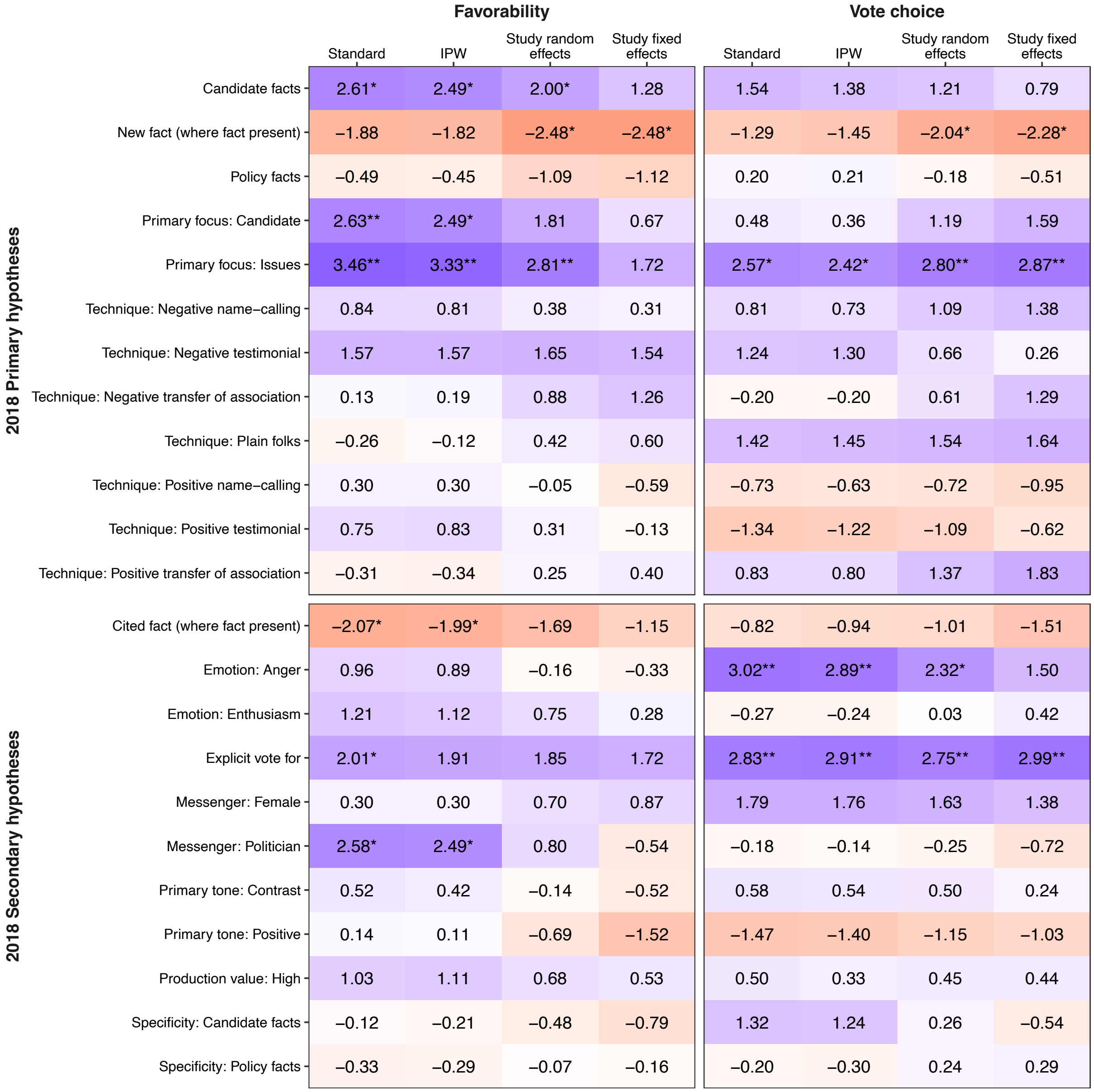
group. The top row in this row are tests comparing the treatment and placebo groups; the bottom row is among treatment groups only. In the 2018 data, the p values of these tests (one per study) show a moderate-sized peak at $p = 0$, reflecting our finding of covariate imbalance discussed above. Note that the covariates in 2018 were measured post-treatment.

Figure OA11: Robustness to attrition.



Notes: Top: Tests of covariate imbalance due to missingness. Top row shows p values by study for a regression $missing * covariate \sim treat$*. Second row shows p values by study for a regression* $missing * covariate \sim content\_id$*, excluding respondents in the placebo group. Bottom: comparison of ATE estimates for each study based on OLS, vs. those inverse-probability-weighted estimates based on a logistic regression model* $missing \sim treat * covariates$*.*

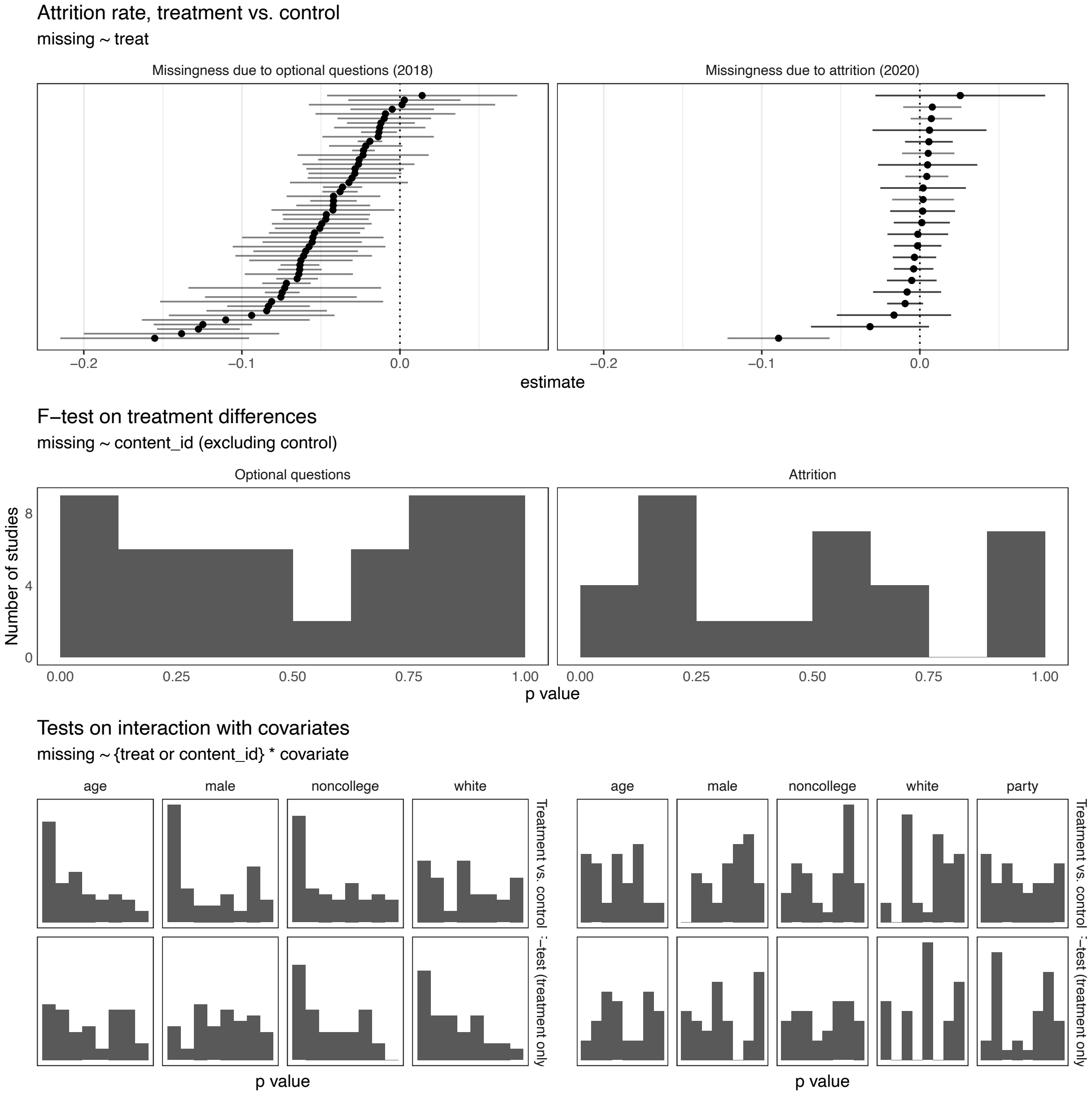
Figure OA12: Comparison of 2018 metaregression results when fit using alternative specifications (compare with Figure 3).



		Favorability				Vote choice			
		Standard	IPW	Study random effects	Study fixed effects	Standard	IPW	Study random effects	Study fixed effects
2018 Primary hypotheses	Candidate facts	2.61*	2.49*	2.00*	1.28	1.54	1.38	1.21	0.79
	New fact (where fact present)	−1.88	−1.82	−2.48*	−2.48*	−1.29	−1.45	−2.04*	−2.28*
	Policy facts	−0.49	−0.45	−1.09	−1.12	0.20	0.21	−0.18	−0.51
	Primary focus: Candidate	2.63**	2.49*	1.81	0.67	0.48	0.36	1.19	1.59
	Primary focus: Issues	3.46**	3.33**	2.81**	1.72	2.57*	2.42*	2.80**	2.87**
	Technique: Negative name−calling	0.84	0.81	0.38	0.31	0.81	0.73	1.09	1.38
	Technique: Negative testimonial	1.57	1.57	1.65	1.54	1.24	1.30	0.66	0.26
	Technique: Negative transfer of association	0.13	0.19	0.88	1.26	−0.20	−0.20	0.61	1.29
	Technique: Plain folks	−0.26	−0.12	0.42	0.60	1.42	1.45	1.54	1.64
	Technique: Positive name−calling	0.30	0.30	−0.05	−0.59	−0.73	−0.63	−0.72	−0.95
	Technique: Positive testimonial	0.75	0.83	0.31	−0.13	−1.34	−1.22	−1.09	−0.62
	Technique: Positive transfer of association	−0.31	−0.34	0.25	0.40	0.83	0.80	1.37	1.83
2018 Secondary hypotheses	Cited fact (where fact present)	−2.07*	−1.99*	−1.69	−1.15	−0.82	−0.94	−1.01	−1.51
	Emotion: Anger	0.96	0.89	−0.16	−0.33	3.02**	2.89**	2.32*	1.50
	Emotion: Enthusiasm	1.21	1.12	0.75	0.28	−0.27	−0.24	0.03	0.42
	Explicit vote for	2.01*	1.91	1.85	1.72	2.83**	2.91**	2.75**	2.99**
	Messenger: Female	0.30	0.30	0.70	0.87	1.79	1.76	1.63	1.38
	Messenger: Politician	2.58*	2.49*	0.80	−0.54	−0.18	−0.14	−0.25	−0.72
	Primary tone: Contrast	0.52	0.42	−0.14	−0.52	0.58	0.54	0.50	0.24
	Primary tone: Positive	0.14	0.11	−0.69	−1.52	−1.47	−1.40	−1.15	−1.03
	Production value: High	1.03	1.11	0.68	0.53	0.50	0.33	0.45	0.44
	Specificity: Candidate facts	−0.12	−0.21	−0.48	−0.79	1.32	1.24	0.26	−0.54
	Specificity: Policy facts	−0.33	−0.29	−0.07	−0.16	−0.20	−0.30	0.24	0.29

Notes: 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).

Figure OA13: Tests for differential attrition in 2018 (left) and 2020 (right).



C Appendix: Returns to experimentation simulation studies

C.1 Mapping survey-based to field estimates

As noted in the main text, immediately measured survey outcomes are likely to overstate the effects of television ads in the field. Our simulations therefore need to scale down in-survey point estimates to likely in-field effects in order to ascertain the likely implications of survey-based results for real-world elections. In particular, for the purposes of our simulations, in order to translate between the survey treatment effects estimated in this study and the real-world impacts of advertising on vote-share, we begin by looking to existing estimates for the cost-effectiveness of campaign TV advertising. For a typical U.S. Senate election, Sides, Vavreck, and Warshaw (2022) estimate that a campaign earns approximately 5 votes per $\$1000$ spent on TV advertising ("5 VPK", or $\$200$ per vote). In our simulations, we then scale this return-on-advertising to vary between ads in direct proportion to their survey-estimated treatments effects – for example, an ad with twice the typical ATE is assumed to yield returns to a campaign of 10 VPK. To simulate the distribution of effects that a campaign produces, we use the μ and τ parameters estimated from Swayable data on vote-choice outcome, averaged over all three datasets (Table 2). We first sample *true* treatment effects from a Normal distribution $N(\mu, \tau)$, and then simulate an experiment by sampling *estimated* treatment effects centered on these true values (with empirical sampling variability). Campaigns are assumed to choose to air the single ad with the largest ATE estimate, and then based on the sampled *true* ATE of this ad we derive quantities such as vote gain. The full set of parameters used in this simulation model is provided in Table 3.

The results from Sides, Vavreck, and Warshaw (2022) are similar to results from Spenkuch and Toniatti (2018). In particular, both studies suggest a ratio between in-field and in-survey effects of very roughly 100, meaning that when an experimental participant watches a persuasive advertisement in a survey experiment, its impact on self-reported vote intention is approximately 100 times larger than the real-world impact that a single exposure to a television advertisement typically has on vote behavior.

Starting with Sides, Vavreck, and Warshaw (2022), the authors estimate ads produce votes at approximately $\$200$ per net vote. To 'work backwards' from this point, we assume that television advertising costs 6c per voter impression (very roughly, 3c per impression in a population for which half of viewers vote). At this rate, the implied treatment effect of television ad exposure on vote share would be approximately 0.015pp for Senate races, and 0.008pp for Presidential races. Compared these values with the average survey treatment effects μ from Table OA2, we calculate the survey-to-field conversion factor to be approximately 100 ($= \frac{1.53}{0.015}$) based on the 2020 Downballot dataset. Spenkuch and Toniatti (2018) provide no estimates for Senate races.

To produce a 'survey-to-field deflation' estimate based instead on the 2020 Presidential dataset, Sides, Vavreck, and Warshaw (2022) estimate the return on Presidential TV advertising to be $\$365$ per vote; in Spenkuch and Toniatti (2018), the same quantity is estimated as $\$170$ per vote. When applying the same method as above relate these estimated returns with Swayable's 2020 Presidential data ($\mu = 1.07$, Table OA2), we estimate the survey-to-field conversion factor to be estimated to be approximately 130 based on Sides, Vavreck, and Warshaw (2022), or to be approximately 60 based on Spenkuch and Toniatti (2018). As an alternative, the latter work provides also its own estimate for the "per-impression" effect of Presidential TV advertising, found to be 0.017pp on vote share per impression per capita[12]. This may be therefore compared directly to the survey ATEs without requiring any further assumptions about the cost of advertising. In this case, the survey-to-field deflation factor is estimated again to be 60 (=$\frac{1.07}{0.017}$).

C.2 Simulating returns to experimentation choices

Campaigns have to choose how much of their advertising budget to invest in experimentation and how to allocate those funds to alternative experimental designs. Here we consider only two design parameters which are the principal decisions campaigns face when determining how much money to invest in ad testing: the number of ads to develop for experimental testing purposes and the total number of subjects to enroll in the experiment.

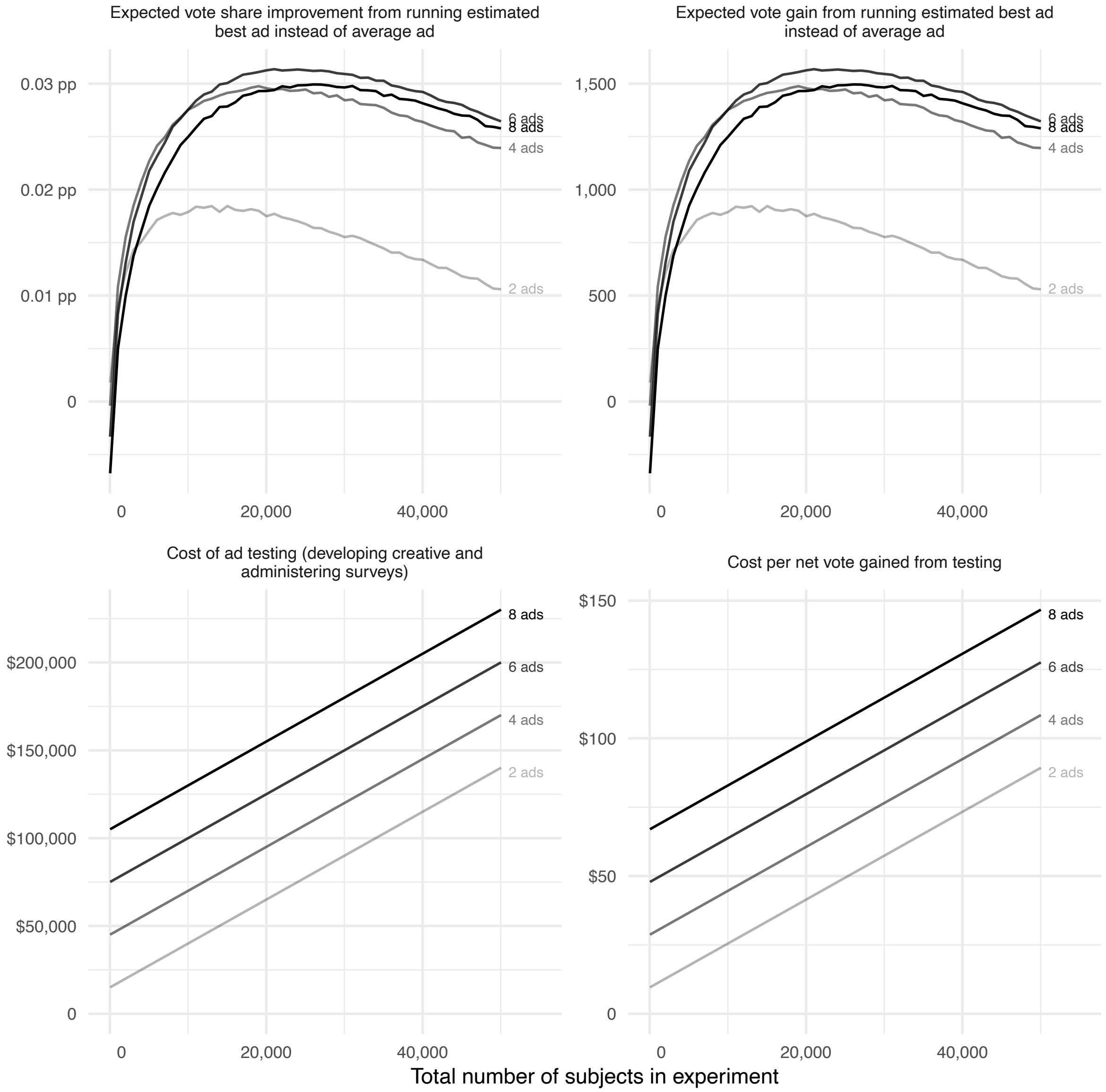
In Figure OA14 we report the results of simulations that illustrate the impacts of these choices on costs and votes gained. Our simulations consider the expected costs and subsequent vote gains from a campaign running experiments on its advertising and then running those advertisements. Using the results of our meta-analysis, we estimated that the treatment effects standard deviation is 0.51 times the average effect (what we referred to as ad variability, or $\frac{\tau}{\mu}$). For the following analysis, we assume a medium sized U.S. Senate campaign, considering $\$1,000,000$ in ad spend for an election where 5,000,000 people vote.

Based on these assumptions, the top left panel of Figure OA14 shows the expected vote share gain a campaign would enjoy by running their best ad instead of an average ad. In these simulations, we simulate a distribution of true treatment effects and the distribution of estimates a campaign would reach, which is a function of both the true treatment effects and sampling variability. The campaign then selects the ad with the highest estimate and exposes voters to it, resulting in a vote share gain equal to the discounted size of the selected ad. As can be seen in the top left panel,

[12] For Presidential election advertising, they estimate that 10 impressions per capita produces a partisan difference in vote shares of 0.304 percentage points. This would translate to a per-impression effect on vote share for a single candidate of 0.017 percentage points (dividing by 10 to translate to impression per capita, by 2 to translate from differences in vote shares between candidates to a single candidate's percentage of the vote). Note that the Spenkuch and Toniatti (2018) estimate already implicitly reflects the fact that not every person who is shown ads votes, and so no further adjustments for voter turnout rates are necessary: "We measure advertising intensity in impressions per capita among voting-aged adults. An impression is defined as one viewer being exposed to one commercial. Our metric of advertising intensity thus corresponds to the number of ads seen by the average adult in a particular DMA" (p. 1993).

our results imply that campaigns can gain an additional 0.03 percentage points in final vote share by testing six ads among 20,000 subjects and running the ad with the highest point estimates.

Figure OA14: Costs and returns to ad experimentation for a typical Senate campaign, assuming ads cost $15,000 to develop and $2.50 per subject to test.



Notes: Lines show simulations where 2, 4, 6, and 8 ads are tested.

Such a gain is politically meaningful. The top right panel of Figure OA14 shows the impact on a candidate's total vote margin such an effect would have in an election where 5,000,000 people vote, such as a typical US Senate race. The expected impact of making six ads, running an experiment with 10,000 subjects, and running the best ad instead of an average ad is an increase of approximately 1,500 net votes.

The bottom left panel shows that this gain comes at a surprisingly small cost. For the sake of our simulations we assume making an additional ad costs $15,000 and that survey experiments on ads' effectiveness cost approximately $2.50 per subject included in the experiment (these rough figures are derived from conversations with political practitioners). Under these assumptions, creating six ads and testing them in an experimental sample of 20,000 subjects would cost $125,000 (above creating just a single ad).

These numbers imply that ad experimentation is an astoundingly compelling investment for campaigns, with a cost per net vote of only $83 in this example. This "cost per vote" is about half the estimated cost per net vote of ad spending itself, and on par with the most cost-effective get out the vote interventions (see Green and Gerber 2019; Spenkuch and Toniatti 2018; Sides, Vavreck, and Warshaw 2022).

C.3 The costs of incorrect beliefs about ad variability

One implication of our simulations is that campaigns can earn more votes to the extent they have accurate beliefs about the extent of ad variability. As shown in Appendix D, this is a very real possibility: without access to the archive we analyze here and only having seen a smaller number of ad experiments, practitioners should have much noisier and, on average, less accurate beliefs about ad variability.

This setting is likely to lead campaigns to make suboptimal resource allocation decisions. Intuitively, if campaigns underestimate ad variability, they will underestimate the returns to experimentation and then underinvest in experimentation. Conversely, if campaigns overestimate the extent of ad variability, they will invest more in experimentation than they should—on average 'wasting' money that should be spent running ads.

In this subsection we illustrate the benefits to campaigns of having correct beliefs about ad variability. In particular, we use the simulations described in the main text to determine how campaigns would optimally allocate resources between ad experimentation and running ads at three budget levels—$500,000, $1,000,000, and $5,000,000—and if they maintained various *subjective beliefs* about ad variability ($\frac{\tau}{\mu}$). For example, in these simulations, campaigns with subjective beliefs that ad variability is tiny would invest nothing in experimentation (as determined by the simulations shown in the main text).However, in the set of simulations in this subsection, we allow campaigns' beliefs to be incorrect. Therefore, although campaigns behave optimally under their beliefs about ad variability, the number of votes their ads actually produce is simulated under the assumption that true $\frac{\tau}{\mu} = 0.51$. For example, a campaign that underestimates the extent of ad variability would not conduct experiments at all, but the true treatment effect of a single ad they make and run would still be drawn from a distribution with the standard deviation $\frac{\tau}{\mu} = 0.51$.

Figure OA15: Votes gained from experimentation if campaigns act optimally under various beliefs about ad variability, if true $\frac{\tau}{\mu} = 0.51$.

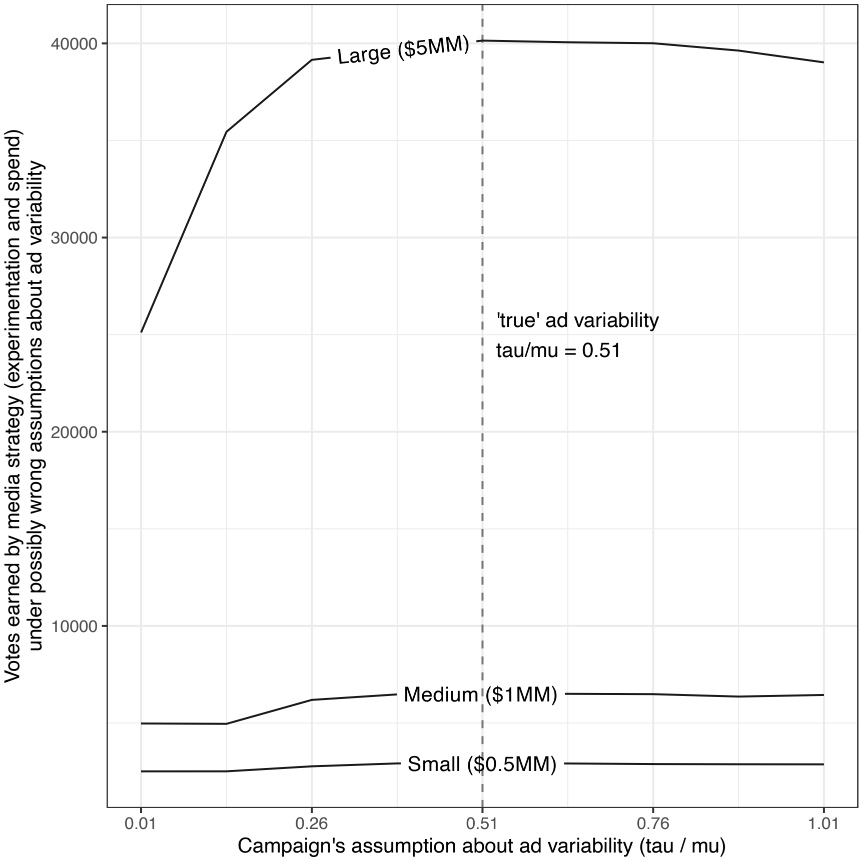


Figure OA15 shows the results. The simulation always assumes that true $\frac{\tau}{\mu} = 0.51$. The horiztonal axis shows campaigns' subjective beliefs about $\frac{\tau}{\mu}$. At the ad variability value of 0.51, campaigns therefore have correct beliefs under the simulation. The vertical axis shows the number of votes that the campaign's ads would produce.

The plot illustrates three key points. First, as can be seen, for all three campaign budget scenarios, the number of votes the campaign gains is maximized when the campaigns have correct beliefs (when campaigns – in the simulation, correctly – believe $\frac{\tau}{\mu} = 0.51$). Campaigns perform less well when their beliefs are incorrect.

Second (and more interesting) is the asymmetric nature of these costs over the distribution of inaccurate beliefs. Campaigns who overestimate $\frac{\tau}{\mu}$ perform slightly less well, but the costs to overestimating $\frac{\tau}{\mu}$ are minimal. This pattern occurs both because the additional funds that campaigns invest in experiments if they believe $\frac{\tau}{\mu}$ is larger are relatively minimal (per Figure 4a, the difference determines how around 3% of the media budget is allocated) and because this funding still does, on average, increase the treatment effects of the selected ad, partially offsetting the decline in left-over funds for media spending. By contrast, the costs of underestimating $\frac{\tau}{\mu}$ are significantly larger. For instance, if campaigns make an error of the same magnitude (0.50) in the negative instead of positive direction, and so believe $\frac{\tau}{\mu} = 0.01$, they do not invest in experimentation at all and perform significantly less well. At values between 0.01 and 0.51, the costs are smaller but remain substantial, and are far greater than the costs of overestimating $\frac{\tau}{\mu}$.

Third and finally, the results show that the above two dynamics are especially acute for larger campaigns. This conclusion follows from the results we showed in the main text that the returns to experimentation are largest for the most well-financed campaigns. Because experiments increase the cost-efficiency of ad spending, they disproportionately benefit well-resourced campaigns. However, as a consequence, the converse is also true: well-resourced campaigns face the largest costs if they underestimate ad variability and therefore fail to experiment.

D Appendix: Contribution of our meta-study

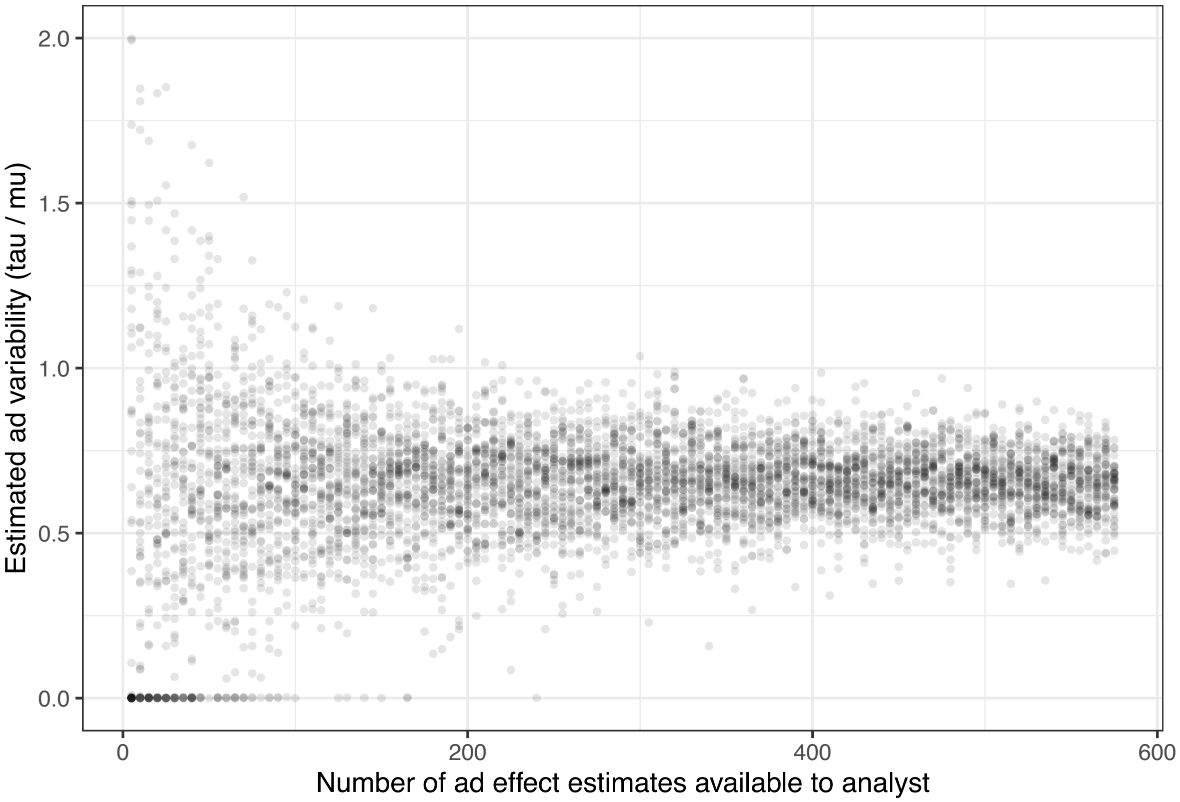
As described above, the returns to experimentation depend crucially on the variability in the effectiveness of ads, as captured by the ratio of the standard deviation of effects to the average effect ($\frac{\tau}{\mu}$). When the variability is low, experimentation is relatively less attractive because all ads perform similarly. When variability is high, experimentation is relatively more attractive. This makes ad variability the key parameter practitioners and scholars must form beliefs about when assessing the value of ad experimentation.

In this appendix, we explore how beliefs about the variability of ad effects might vary depending on how many ad effects are available to the analyst. This speaks to the contribution of our study because it shows that individuals (e.g., individual scholars, campaigns, or campaign consultants) who have seen a smaller number of experimental results than we have in our dataset would have much less accurate beliefs about the extent to which advertising effects vary than we offer. In other words, we show that access to the conclusions from our dataset will allow many scholars and practitioners to form much more precise beliefs about ad variability than they would have been able to otherwise.

In particular, Figure OA16 shows the results of a simulation study in which we estimate the variability of advertisement effects ($\frac{\tau}{\mu}$) from differently-sized subsets of the effects of our ads on vote choice. The horizontal axis describes the number of advertisements available to the analyst, from a low of five ads to a high of 575 ads, the total number of average effects on vote choice in our meta-study. The vertical axis displays the estimate of the standard deviation of effects from the corresponding meta regression. For each number of ads, we simulate the resulting estimate of $\frac{\tau}{\mu}$ 50 times, sampling the appropriate number of ads effects from our dataset with replacement.

The figure shows that when analysts have seen the effects of only few ads, they might have very heterogeneous beliefs about the extent to which ad effects vary, but as the number of ads studies increases, beliefs about variability sharpen up quite a bit.

Figure OA16: Distribution of tau estimates, depending on number ad effect estimates available to analysts



This simulation underlines the contribution of our meta-study. Understanding the distribution of ad effects is difficult, even when an analyst has access to dozens of estimates.

References for Appendices

Green, Donald P., and Alan S. Gerber. 2019. *Get Out The Vote: How to Increase Voter Turnout*. 3rd ed. Washington, DC: Brookings Institution Press.

Sides, John, Lynn Vavreck, and Christopher Warshaw. 2022. “The Effect of Television Advertising in United States Elections.” *American Political Science Review* 116 (2): 702–718.

Spenkuch, Jörg L, and David Toniatti. 2018. “Political Advertising and Election Results.” *The Quarterly Journal of Economics* 133 (4): 1981–2036.