Do Voters Prefer Women and Young Candidates?

Re-evaluating Evidence from Conjoint Experiments

Scott F Abramson*   Bahar Zafer†

August 17, 2026

Abstract

Conjoint experiments are widely used to study whether voters discriminate against candidates on the basis of characteristics such as gender and age. Recent meta-analyses interpret this literature as evidence against demand-side explanations for the under-representation of women and younger candidates in elected office. We reassess these claims by examining the finite-sample sensitivity of candidate-choice conjoint estimates. Building on recent advances in influence-function methods, we estimate the smallest share of respondents and experimentally generated electoral contests whose removal changes the sign of study-level and meta-analytic AMCEs. Across experiments in prominent meta-analyses of gender and age, we find that many estimates can be reversed by dropping small fractions of respondents or contests. These sensitivities also propagate to the meta-analytic level, where pooled estimates change sign after similarly small perturbations. The results show that conjoint-based claims about voter preferences can depend heavily on both respondent sampling and the random sampling of candidate contests implied by experimental designs.

Keywords: Gender bias, candidate-choice experiments, discrimination, sensitivity analysis

*Associate Professor, Department of Political Science, University of Rochester, email: sabramso@ur.rochester.edu

†PhD Candidate, Department of Political Science, University of Rochester, email: bzafer@ur.rochester.edu

Conjoint designs are widely used in political science to estimate the effects of candidate characteristics, such as gender or age, on voters' choices. Recent meta-analyses of candidate-choice experiments on gender (Schwarz and Coppock, 2022) and age (Eshima and Smith, 2022), both published in *The Journal of Politics*, summarize the existing evidence as weighing against "demand-side" explanations for the over-representation of men and older candidates in politics. That is, they interpret the accumulated conjoint evidence as showing little voter discrimination against women or younger candidates. In this note, we use tools from the econometric analysis of finite-sample robustness to examine how stable these conclusions are to the removal of small subsets of observations. We find that, at both the individual-study and meta-analytic levels, the resulting conclusions are fragile.

The Average Marginal Component Effect (AMCE) — the causal estimand targeted in most conjoint experiments — averages in two distinct ways: across respondents and across the set of electoral contests generated by the researcher's randomization scheme (Abramson et al., 2024; Bansak et al., 2023). We extend the Approximate Maximum Influence Perturbation of Giordano, Meager and Broderick (2026*a*,*b*) to evaluate the sensitivity of conjoint estimates to both forms of averaging. Specifically, we ask how many *respondents* and *electoral contests* must be removed, respectively, to change the sign of a given AMCE. We then extend the same logic to meta-analysis, first allowing removals to be allocated freely across studies and then imposing study-level constraints that limit the share of observations removed from any single experiment.

The results reveal substantial sensitivity. Approximately one-third of the experiments in the meta-analyses of Schwarz and Coppock (2022) and Eshima and Smith (2022) require removing less than 2% of respondents to reverse the sign of the estimated gender and age effects, respectively. Contest-level perturbations are similarly consequential: about 60% of the experiments analyzed by both Schwarz and Coppock and Eshima and Smith require dropping less than 2% of electoral contests to reverse the relevant effect. At the meta-analytic level, the sign of the Schwarz and Coppock estimate can be reversed by dropping only 0.72% of respondents when removals are unconstrained across studies. When removals are capped within each experiment, the estimate still changes sign with a study-level cap of 2.7%, corresponding to the removal of 1.65% of respondents overall. The meta-analytic average can likewise be reversed by dropping 0.49% of electoral contests without cross-study constraints and 1.71% with such constraints at 2.3%.

These findings suggest that interpretations of the AMCE are highly sensitive to small perturbations in the sampling process. This sensitivity arises both from the sampling of respondents from the target population and from the experimental sampling of candidate profiles from the distribution of potential electoral contests. Even aggregate findings from candidate-choice conjoint designs should therefore be interpreted with considerable caution.

Maximum Individual Influence Score

Our goal is to assess how sensitive the mean-preference interpretation of the AMCE is to dropping a small number of respondents.[1] We build on Giordano, Meager and Broderick (2026*a*,*b*), who provide a computationally feasible method for approximating the smallest set of observations required to change the sign, or significance, of an estimated treatment effect. Because conjoint designs typically elicit multiple choices from each respondent, we aggregate observation-level influence scores to the respondent level and estimate the smallest set of respondents whose removal reverses a target AMCE. For each study, we first replicate the target AMCE by estimating

$$y_{ijc} = \beta_0 + \beta_1 T_{ijc} + X_{ijc}\beta_2 + \epsilon_{ijc},$$

where y_{ijc} indicates whether respondent i chose profile c in task j, T_{ijc} is an indicator that profile c contains the focal attribute level t_1 rather than the baseline level t_0, and X_{ijc} contains the remaining randomized profile attributes. In a binary conjoint task, $c \in \{1, 2\}$, and $\hat{\beta}_1 = \hat{\pi}(t_1, t_0)$ estimates the AMCE of attribute level t_1 relative to t_0.[2]

We then ask what minimum share of respondents, $r \in [0, 1]$, must be removed to induce a perturbation large enough to reverse the sign of $\hat{\pi}(t_1, t_0)$. For sign reversal, the required perturbation is $\Delta^{+/-} = |\hat{\pi}(t_1, t_0)|$. For each observation, indexed by respondent i, task j, and profile c, let ψ_{ijc} denote the influence score computed using the approximation in Giordano, Meager and Broderick (2026*a*,*b*). We aggregate these to the respondent level, $\psi_i = \sum_{j=1}^{J_i} \sum_{c=1}^{2} \psi_{ijc}$. Suppose $\hat{\pi}(t_1, t_0) > 0$; the negative case is symmetric. Order respondents so that

$$\psi_{i_{(1)}} \geq \psi_{i_{(2)}} \geq \cdots \geq \psi_{i_{(|N|)}}.$$

Dropping respondents in this order gives the influence-based upper bound

$$U = \min\left\{ n \in \{1, \ldots, |N|\} : \sum_{k=1}^{n} \psi_{i_{(k)}} > \Delta^{+/-} \right\}.$$

[1] In Appendix Section A1 we develop the analogous approach to contests.

[2] We use the sampling weights from the original study when applicable and cluster standard errors by respondent.

We then re-estimate the AMCE after removing $D_U = \{i_{(1)}, \ldots, i_{(U)}\}$ to verify that the sign reverses.

Because the influence scores are approximate, this procedure can overstate the number of respondents required. We therefore perform a correction step. For each $m < U$, define the nested drop set $I_m = \{i_{(1)}, \ldots, i_{(m)}\}$. We re-estimate the AMCE after dropping each I_m and define

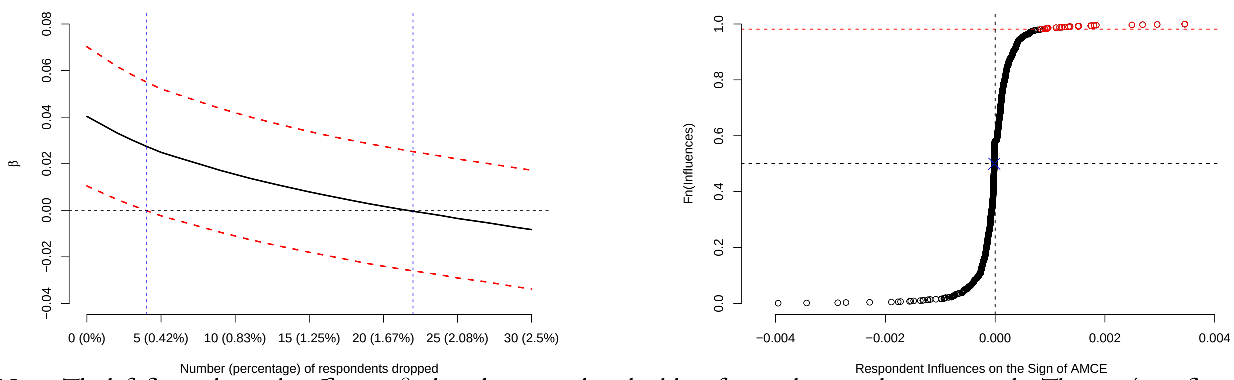
$$m^* = \min\{m < U : \text{sign}(\hat{\pi}_{-I_m}(t_1, t_0)) \neq \text{sign}(\hat{\pi}(t_1, t_0))\}.$$

The estimated most influential set is therefore[3] $I^* = \{i_{(1)}, \ldots, i_{(m^*)}\}$, and the estimated respondent-drop proportion is $\hat{r} = m^*/|N|$.

An Example of a Single Study

To illustrate the details of our application with an example, we consider the YouGov experiment in the paper of Kirkland and Coppock (2018). This candidate choice experiment has 1200 respondents and 11,432 observations. We identically replicate their results with respect to the AMCE of female, a point estimate of .04 with a 95% confidence interval of $[0.010, 0.070]$. We exploit our method to characterize the minimal number of individual respondents necessary to flip the sign of the AMCE. The left panel of Figure 1 plots the number of individuals our method drops on the x-axis against the estimated AMCE of *female* on the y-axis. We see that by dropping four individuals (0.3%) of the sample, the sign of the AMCE becomes statistically indistinguishable from zero. Dropping a total of 22 individuals (1.8%) from the sample turns the

Figure 1: Influence of Respondents on the AMCE of Female Candidate in Kirkland and Coppock (2018) YouGov Experiment



Note: The left figure draws the effect on β when dropping these highly influential respondents iteratively. The sign (significance) of β reverses when 22 (four) respondents are dropped. The right figure shows the empirical CDF of respondent influences on the AMCE of female candidate. The median respondent marked in blue has a negative influence score while being very close to zero. 22/1200 respondents (red) are dropped to reverse the sign.

[3]This correction is necessary because aggregation can compound approximation error in the observation-level influence scores. In our applications, the correction frequently reduces the estimated minimum number of respondents required to reverse the target AMCE.

point estimate negative. In the right panel of Figure 1 we plot the empirical CDF of the individual influence scores derived from this exercise. The 22 largest positive scores (highlighted in red) are those sufficient to change the sign.

Sensitivity of Meta-analysis to Removing Respondents

We next extend the respondent-level perturbation exercise to the meta-analytic estimand. The goal is to approximate the smallest set of respondents, drawn across the set of included conjoint experiments, whose removal changes the substantive interpretation of the pooled meta-analytic estimate.

Let $s = 1, \dots, S$ index experiments. For each replicated experiment s, we estimate the target AMCE, $\hat{\pi}_s(t_1, t_0)$, and its estimated standard error, $\widehat{\text{se}}(\hat{\pi}_s(t_1, t_0))$, using the same conjoint specification as in the individual-study analysis. We then compute respondent-level influence scores for the target AMCE within each experiment. Let ψ_{is} denote the aggregate influence of respondent i in experiment s, obtained by summing the observation-level influence scores over all conjoint tasks and profiles completed by that respondent:

$$\psi_{is} = \sum_{j=1}^{J_{is}} \sum_{c=1}^{2} \psi_{ijcs}.$$

The meta-analytic estimate is a weighted average of the study-specific AMCEs,

$$\hat{\Pi} = \sum_{s=1}^{S} w_s \hat{\pi}_s(t_1, t_0),$$

where w_s denotes the meta-analytic weight assigned to experiment s. The corresponding standard error is computed using the same meta-analytic procedure used in the original analysis.

To construct the perturbation path, we pool the respondent-level influence scores from all replicated experiments into a single data set. Suppose that the original meta-analytic estimate is positive; the negative case is symmetric. We order respondents by the direction and magnitude of their influence on the target AMCE:

$$\psi_{i_{(1)} s_{(1)}} \geq \psi_{i_{(2)} s_{(2)}} \geq \cdots \geq \psi_{i_{(M)} s_{(M)}}, \qquad M = \sum_{s=1}^{S} N_s, \qquad \psi_{i_{(k)} s_{(k)}} = \psi_{(k)}$$

For each $m = 1, \dots, M$, define the nested respondent-drop set $I_m = \{(i_{(1)}, s_{(1)}), \dots, (i_{(m)}, s_{(m)})\}$. At each step m, we remove the respondents in I_m from their respective experiments. We then re-estimate each affected conjoint experiment and update its AMCE estimate and standard error. Experiments from which no respondent has been removed retain their original estimates. Let $\hat{\pi}_{s,-I_m}(t_1, t_0)$ and $\widehat{\text{se}}(\hat{\pi}_{s,-I_m}(t_1, t_0))$ denote the updated study-specific coefficient and standard error after removing the respondents in I_m.

Using these updated estimates, we re-estimate the meta-analysis at each step:

$$\hat{\Pi}_{-I_m} = \sum_{s=1}^{S} w_{s,-I_m} \hat{\pi}_{s,-I_m}(t_1, t_0),$$

where $w_{s,-I_m}$ denotes the updated meta-analytic weight. We then define the smallest respondent set that reverses the sign of the meta-analytic estimate as

$$m^*_{meta} = \min \left\{ m : \text{sign}\left(\hat{\Pi}_{-I_m}\right) \neq \text{sign}\left(\hat{\Pi}\right) \right\}, \qquad \hat{r}_{meta} = \frac{m^*_{meta}}{\sum_{s=1}^{S} N_s}.$$

Constrained Perturbation

The unconstrained perturbation exercise identifies the globally most influential respondents across all experiments. Although this provides a natural benchmark, it may concentrate removals in a small number of studies while leaving other studies unchanged. To assess whether the sensitivity of the meta-analytic result is driven by a small number of experiments, we repeat the analysis subject to an experiment-level constraint on the share of respondents that may be removed.

Let $\tau \in (0, 1)$ denote the maximum permissible respondent-drop share within any experiment. For each experiment s, the constrained perturbation set must satisfy $|I_m \cap s| \leq \tau N_s$, where $|I_m \cap s|$ is the number of dropped respondents from experiment s. We therefore construct a constrained ordering of respondents by proceeding through the pooled influence ranking but excluding any respondent from experiment s once that experiment has reached its threshold τN_s.

For each value of τ, we then repeat the iterative procedure above: respondents are removed according to the constrained influence ranking; the affected conjoint experiments are re-estimated; and the meta-analysis is updated using the revised study-level coefficients and standard errors.

Results

We replicate the random-effects meta-analyses in Schwarz and Coppock (2022) and Eshima and Smith (2022).[4] Of the 70 distinct experiments across the 67 studies in the gender meta-analysis, we were able to obtain replication materials for 54. Of the 18 experiments in the age meta-analysis, we were able to obtain replication materials as shared by Eshima and Smith (2022)[5]. In the main text, we report results using all 70 and 18 experiments, respectively. When re-estimating the meta-analytic average effect, we treat experiments without replication materials as contributing no dropped observations. For the contest-level results, we calculate the total proportion of dropped contests by imputing, for experiments without replication data, the expected number of electoral contests implied by each experiment's design as described in the original paper.[6] In Appendix Section A5, we replicate the full exercise using only the subset of experiments for which replication materials were available.

Sensitivity of Individual Studies

See Section A4 of the appendix where we report study-level results for every individual study with full replication data in both meta-analyses. Tables A3 and A4 show that the individual-study results are highly sensitive to small perturbations in both respondent samples and experimentally generated contests. In the gender experiments, 18 of 54 replicated studies, or 33.3%, require dropping less than 2% of respondents to reverse the sign of the AMCE of female versus male candidates; 39 of 54, or 72.2%, require dropping less than 5%. Contest-level perturbations are even more pronounced: among the 42 gender experiments for which contest-level analysis is possible, 24, or 57.1%, reverse sign after removing less than 2% of contests, and 36, or 85.7%, reverse after removing less than 5%. The age experiments show a similar pattern. For the AMCE of the oldest versus youngest candidate, 6 of 18 experiments, or 33.3%, reverse sign after removing less than 2% of respondents, while 11 of 18, or 61.1%, reverse after removing less than 5%. At the contest level, 10 of the 17 analyzable age experiments, or 58.8%, reverse after removing less than 2% of contests, and 13 of 17, or 76.5%, reverse after removing less than 5%. Taken together, these results indicate that the sensitivity documented in the pooled meta-analyses is already present in a large share of the underlying experiments: many study-level AMCEs depend on a small fraction of either sampled respondents or sampled candidate contests.

[4] See Table A1 and Table A2.

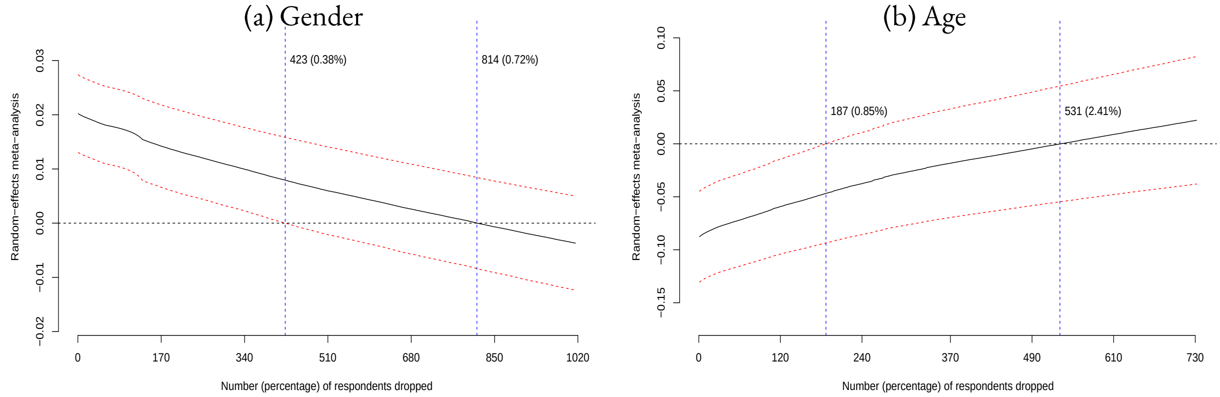
[5] The meta-analysis in Eshima and Smith (2022) has 16 replications. We have 18 since we replicate MTurk and YouGov experiments in Kirkland and Coppock (2018) separately for partisan and non-partisan contests. For all experiments except one, whose replication does not contain a variable indicating paired profiles in each contest, we are able to conduct contest-level analysis.

[6] For details, see Section A2.

Sensitivity of Meta-analysis to the Removal of Individuals

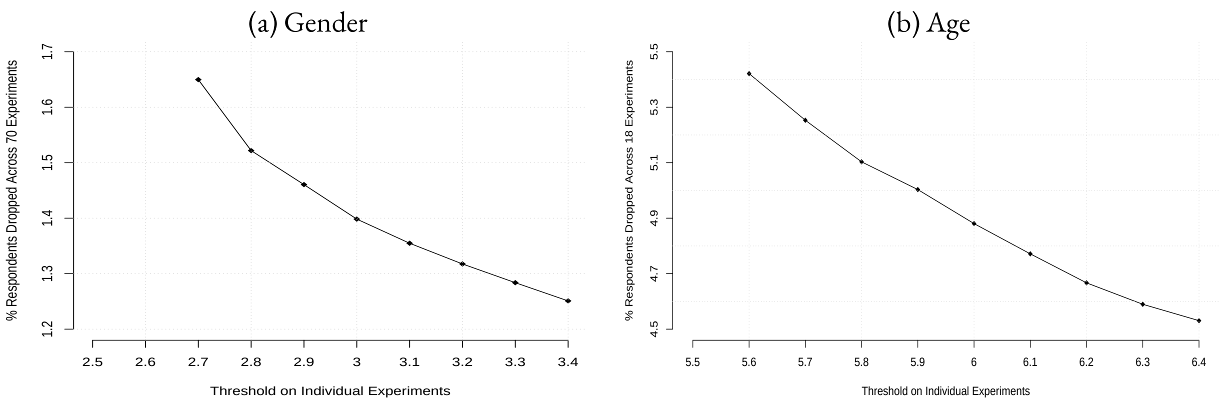
To start, we give results describing the sensitivity of both meta-analyses to the removal of individual respondents. Results from the unconstrained exercise (placing no cap on the proportion of respondents dropped from any single experiment) are given in Figure 2. In the left panel we give results for the gender meta analysis and in the right panel we give results for age[7] For gender, it takes dropping just 423 out of 112561 respondents — just 0.38% — to alter the significance (make it null) of the meta-analytic average. To alter the sign dropping an additional 389 observations — 0.72% of the total — changes the sign. We obtain similar results for the meta-analytic average effect of age: here, we find that dropping 187 individuals — 0.85% of the total — changes the significance; dropping 531 individuals — a total of 2.41% of all respondents — changes the sign of the meta-analytic average effect.

Figure 2: Effect of Dropping Respondents on Meta-analysis Estimates



Note: Figures show how the effects change as we remove the most influential respondent iteratively from the meta-analysis of AMCE(Female, Male) (left) and AMCE(Oldest, Youngest) (right).

Figure 3: Change in Percentage of Respondents Dropped to Reverse the AMCE Sign by Threshold



Note: Figures show how % respondents dropped to reverse the effect changes as we numerically search for the minimum cap on the maximum % for individual experiments in the meta-analysis of gender (left) and age (right).

[7] As in Eshima and Smith (2022) we treat the baseline category for the AMCE as the youngest and second oldest categories, respectively. In the main text we give results using the former. Results using the latter are very similar and are presented in Appendix Section A8.

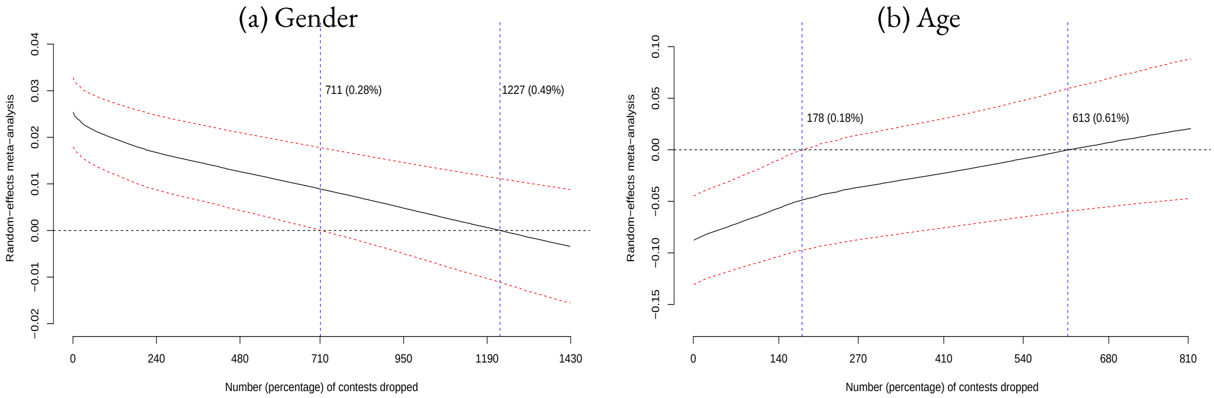
This procedure, of course, imposes no restriction on which experiments supply the respondents selected for removal. Figure 3 therefore reports constrained estimates that limit the maximum share of respondents that can be dropped from any single experiment. The x-axis shows this imposed cap, while the y-axis shows the percentage of all respondents that must be removed to change the sign of each meta-analytic average effect. For gender, the smallest per-experiment cap under which the sign can still be reversed is 2.7%; at this cap, removing 1.65% of all respondents flips the AMCE. For age, the smallest feasible cap is 5.6%, under which reversing the sign requires dropping 5.4% of all respondents.

Sensitivity of Meta-analysis to the Removal of Contests

We repeat the exercise, now focusing on the sensitivity of each meta-analysis to the removal of electoral contests. The unconstrained results are presented in Figure 4. The left panel reports results for the gender meta-analysis. Changing the statistical significance of the meta-analytic estimate requires dropping 711 contests, or 0.28% of all contests. Changing its sign requires dropping only 516 additional contests, or 0.49% of the total. The results for the age meta-analysis, shown in the right panel, are similar. In this case, changing statistical significance requires dropping 178 contests, or 0.18% of the total, while changing the sign requires dropping 616 contests, or 0.61% of the total.

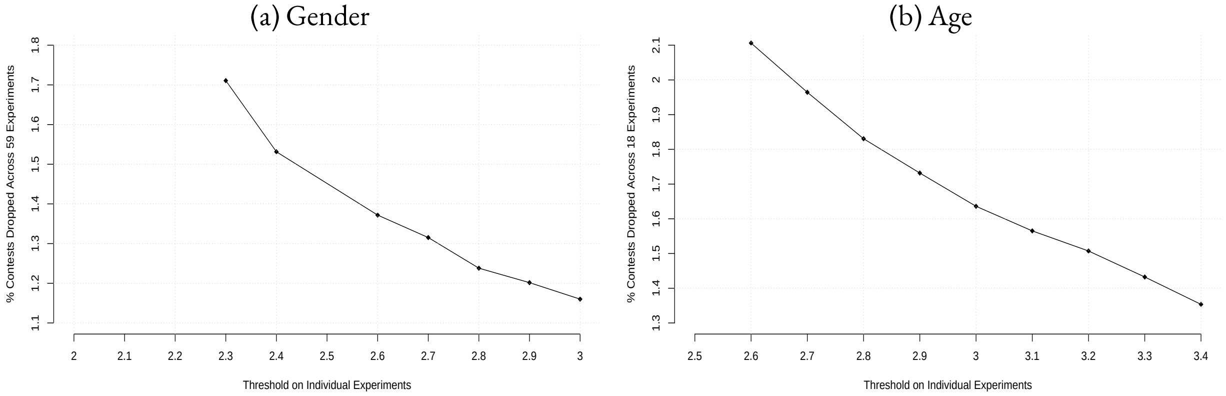
The constrained results, reported in Figure 5, tell a similar story. For gender, the smallest per-experiment cap under which we can still reverse the sign of the meta-analytic average is 2.3%, corresponding to the removal of 1.71% of all contests. For age, the analogous cap is 2.6%, which yields a total contest-removal rate of 2.1%. These percentages require some care in interpretation. Because the experiments use different randomization schemes, and therefore differ in the dimensionality of their factorial designs, removing a contest

Figure 4: Effect of Dropping Contests on Meta-analysis Estimates



Note: Figures show how the effects change as we remove the most influential contest iteratively from the meta-analysis of AMCE(Female, Male) (left) and AMCE(Oldest, Youngest) (right).

Figure 5: Change in Percentage of Contests Dropped to Reverse the AMCE Sign by Threshold



Note: Figures show how % contests dropped to reverse the effect changes as we numerically search for the minimum cap on the maximum % for individual experiments in the meta-analysis of gender (left) and age (right).

mechanically removes different numbers of data-points across studies. As a result, the per-experiment cap is comparatively large and still tends to favor removing observations from lower-dimensional experiments. We return to this issue in Appendix A6, where we examine how experiment dimensionality is associated with both individual-study and meta-analytic sensitivity.[8]

Discussion

The results we present in this note point to a broader interpretive problem for conjoint designs. The AMCE is often read as a stable summary of voter preferences, but the estimates we report show that this interpretation can depend heavily on two forms of sampling: the sampling of individuals from the target population and the experimental sampling of candidate contests from the researcher's randomization scheme. Because conjoint estimates average over both respondents and the realized distribution of profile comparisons, small changes in either sampled respondents or sampled contests can meaningfully alter the estimated effect, including its sign. This is not only a study-level concern. The same instability propagates to meta-analysis, where pooled estimates inherit the sampling dependence of the underlying experiments and can be reversed by removing relatively small numbers of influential respondents or contests across studies. The implication is not that conjoint designs are uninformative, but that their preference-based interpretation requires greater caution: when conclusions depend on a small subset of respondents or a small subset of randomly generated electoral contests, the AMCE may be better understood as a sample-dependent summary of a

[8] In the same section, we also examine how the number of experimental subjects is associated with both individual-study and meta-analytic sensitivity. In Appendix A7, we develop and apply, to the set of gender experiments, an approach for examining how individual covariates are associated with respondent-level influence scores. We show for the AMCE of female vs male that men and Republicans have significantly more negative influence scores than women and Democrats.

particular design realization than as direct evidence for or against broad claims about voter tastes.

References

Abramson, Scott F, Korhan Kocak, Asya Magazinnik and Anton Strezhnev. 2024. "Aggregation, interpretation, and estimation of preferences in conjoint experiments." *SocArXiv https://osf.io/preprints/socarxiv/xjre9* .

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2023. "Using conjoint experiments to analyze election outcomes: The essential role of the average marginal component effect." *Political Analysis* 31(4):500–518.

Eshima, Shusei and Daniel M. Smith. 2022. "Just a Number? Voter Evaluations of Age in Candidate-Choice Experiments." *The Journal of Politics* 84(3):1856–1861.

Giordano, Ryan, Rachael Meager and Tamara Broderick. 2026*a*. "An automatic finite-sample robustness metric: when can dropping a little data change conclusions? Part I: definitions and experiments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 384(2321):20250001.

Giordano, Ryan, Rachael Meager and Tamara Broderick. 2026*b*. "An automatic finite-sample robustness metric: when can dropping a little data change conclusions? Part II: theory and intuition." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 384(2321):20240614.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Kirkland, Patricia and Alexander Coppock. 2018. "Candidate Choice Without Party Labels:: New Insights from Conjoint Survey Experiments." *Political Behavior* 40:571–591.

Schwarz, Susanne and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84(2):655–668.

Appendix

Contents

A1 Contest Sensitivity

Consider the conjoint estimation that we used to illustrate our respondent sensitivity analysis:

$$y_{ijc} = \beta_0 + \beta_1 T_{ijc} + X_{ijc}\beta_2 + \epsilon_{ijc} \tag{1}$$

where y_{ijc} is an indicator of whether respondent i chooses profile c in question j. T_{ijc} indicates whether profile c contains the focal attribute level t_1 rather than the baseline level t_0. X_{ijc} contains the remaining randomized profile attributes. $c \in \{1, 2\}$, since there are two profiles in each question. $\hat{\beta}_1 = \hat{\pi}(t_1, t_0)$ is the AMCE of t_1 relative to t_0. We estimate the minimum proportion of contests we must drop in order to achieve a perturbation at the size of $\Delta = |\hat{\pi}(t_1, t_0)|$ to reverse the sign of $\hat{\beta}_1$.

We define a profile as a unique combination of all L attributes in the experiment. Let x_{ijc} be a vector of the attributes of profile c in question j answered by respondent i.

$$\mathbf{x}_{ijc} = (T_{ijc}, x_{ijc}^{(1)}, x_{ijc}^{(2)}, ..., , x_{ijc}^{(L-1)}) \tag{2}$$

We define the function ϕ that maps the attribute vector to a unique profile identifier.

$$\mathbf{p}_{ijc} = \phi(\mathbf{x}_{ijc}) \tag{3}$$

The set of unique profiles observed in the experiment is

$$\mathcal{P}_{unique} = \{\mathbf{p} \in \mathbb{P} \mid \exists (i, j, c) \text{ such that } \mathbf{x}_{ijc} = p\} \tag{4}$$

where $\mathbb{P} = \{t_1, t_0\} \times A_1 \times A_2 \times ... \times A_{L-1}$ is the space of all possible profiles and A_l is a set of possible levels for attribute l.

We define a contest as an unordered pair of two profiles.

$$\mathbf{c}_{ab} = \{\mathbf{p}_a, \mathbf{p}_b\} \tag{5}$$

where $\mathbf{p}_a, \mathbf{p}_b \in \mathcal{P}_{unique}$ and $\mathbf{p}_a \neq \mathbf{p}_b$. Then, the unique set of contests observed in the experiment is

$$\mathcal{C}_{unique} = \{\{\mathbf{p}_a, \mathbf{p}_b\} \mid \exists (i, j) \text{ such that } \{\mathbf{x}_{ij1}, \mathbf{x}_{ij2}\} = \{\mathbf{p}_a, \mathbf{p}_b\}\} \tag{6}$$

We compute the influence score of a data point ψ_{ijc} using the package provided by Giordano, Meager and Broderick (2026*a*,*b*). The influence score of a question j answered by respondent i is $\psi_{ij} = \psi_{ij1} + \psi_{ij2}$. We define the influence score of a unique contest $\mathbf{c} \in \mathcal{C}_{unique}$ as the sum of influence scores of questions that ask $\mathbf{c}$.

$$\psi_{\mathbf{c}} = \sum_{i=1}^{N} \sum_{j=1}^{J_i} \mathbb{1}(\{\mathbf{x}_{ij1}, \mathbf{x}_{ij2}\} = \mathbf{c}) \times \psi_{ij} \tag{7}$$

As before, we rank contests based on their influence scores, $\mathbf{c}_{(1)} > \mathbf{c}_{(2)} > ... > \mathbf{c}_{(|\mathcal{C}_{unique}|)}$, such that $\mathbf{c}_{(k)}$ denotes the contest $\mathbf{c}$ with the kth largest influence score, $R(\mathbf{c}) = k$.

Then, the minimum number of contests that we need to drop to reverse the sign of $\hat{\pi}(female, male)$ is

$$U_{\mathbf{c}} = min\left\{k \in \{1, 2, ..., |\mathcal{C}_{unique}|\} \middle| \sum_{z=1}^{k} \psi_{\mathbf{c}_{(z)}} > \Delta, \mathbf{c} \in \mathcal{C}_{unique}\right\} \tag{8}$$

Given $U_{\mathbf{c}}$, the set of contests that we drop is

$$D_{\mathbf{c}} = \{\mathbf{c} \in \mathcal{C}_{unique} \mid R(\mathbf{c}) \leq U_{\mathbf{c}}\} \tag{9}$$

We apply our numerical correction step as before and get the set of the most influential contests to reverse the sign of the AMCE of female.

$$D^*_{\mathbf{c}} = \{\mathbf{c} \in \mathcal{C}_{unique} \mid R(\mathbf{c}) \leq U^*_c\} \tag{10}$$

where $D^*_{\mathbf{c}} \subseteq D_{\mathbf{c}}$ and $U^*_{\mathbf{c}} \leq U_{\mathbf{c}}$. Thus, the minimum share of contests that can be dropped to reverse the sign of $\hat{\pi}$ is $\frac{|D^*_{\mathbf{c}}|}{|\mathcal{C}_{unique}|}$.

A2 Calculation of Expected Number of Contests

When we do not have the replication data for a conjoint experiment, we calculate the expected number of unique contests observed in the experiment using the number of respondents, N, the number of questions answered by each respondent, J, and the total number of unique candidate profiles, K.

Let C denote the total possible number of unique two-candidate matchups (contests) that can be formed from K profiles. Assuming candidate order does not matter and candidates cannot run against themselves,

C is given by the combination formula:

$$C = \binom{K}{2} = \frac{K(K-1)}{2}$$

Let Q_c indicate whether a unique contest, c, is drawn and X be the observed number of contests in the experiment.

$$X = Q_1 + Q_2 + \cdots + Q_C, \quad Q_c = \begin{cases} 1, & \text{if } c \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

When we have the replication data for an experiment, we observe whether a contest c is in the data and compute X directly. Otherwise, we calculate the expectation of X.

Let T denote the total number of experimental trials (matchups generated). Since N respondents answer J questions, we have $T = N \times J$. Since both profiles are randomized in conjoint designs, a unique contest, c, has the probability of being drawn in any single trial, $P(\text{c is drawn}) = \frac{1}{C}$ and across all trials, $P(\text{c is never drawn}) = (1 - \frac{1}{C})^T$. Then, the expected value of X:

$$\begin{aligned} E(X) &= P(Q_1 = 1) + P(Q_2 = 1) + ... + P(Q_C = 1) \\ &= C - (P(Q_1 = 0) + P(Q_2 = 0) + ... + P(Q_C = 0)) \\ &= C - C\left(1 - \frac{1}{C}\right)^T \\ &= C\left(1 - \left(\frac{C-1}{C}\right)^T\right) \end{aligned}$$

A3 Sensitivity of Meta-analysis Effect

Meta-analysis of AMCE(Female, Male)

Table A1 shows the results of our replication of the random-effects meta-analysis in Schwarz and Coppock (2022). With any subset of experiments pertaining to our respondent- or contest-level analysis, we estimate a random-effects meta-analysis coefficient around 0.02. Hence, our procedure closely replicates the findings of Schwarz and Coppock (2022) by finding a little larger effect. The meta-analysis coefficients with the experiments that we are able to analyze deviate from the coefficient estimated by Schwarz and Coppock (2022) by approximately 0.006. This means that we need to achieve a perturbation of an even bigger size to reverse the sign of the effect compared to the ideal scenario in which we would be able to analyze all experiments.

In our respondent-level analysis, we included all experiments in the meta-analysis of Schwarz and Coppock (2022)[9]. In cases where we were not able to find the replication data[10] or replicate the AMCE in the original paper or Schwarz and Coppock (2022)[11], we take the coefficients, standard errors, and number of respondents from Schwarz and Coppock (2022). For five experiments, we were unable to find the reference article[12] or the experiment[13], and we also take the number of respondents from Schwarz and Coppock

Table A1: Replication of Random-effects Meta-analysis of AMCE(Female, Male)

	Estimate	Std.error	Conf.low	Conf.high	# Experiments
Schwarz and Coppock (2022)	0.0184	0.0036	0.0115	0.0254	67
Respondent analysis					
All Experiments	0.0202	0.0037	0.0131	0.0274	70
Analyzed Experiments	0.0242	0.0041	0.0162	0.0322	54
Contest analysis					
All Experiments	0.0254	0.0038	0.018	0.0328	59
Analyzed Experiments	0.0243	0.0045	0.0156	0.0331	42

Note: Table shows our replications of the meta-analysis of AMCE(Female, Male) using the experiments included in respondent (Row 2 and 3) and contest (Row 4 and 5) analysis along with the original meta-analysis estimates in Schwarz and Coppock (2022) (Row 1).

[9]We excluded two experiments in Vivyan and Wagner (2015) to avoid double-counting since they are the same experiments also published in Campbell et al. (2019).

[10]Fox and Smith (1998), Hobolt and Rodon (2020), Piliavin (1987), Sen (2017), Sigelman and Sigelman (1982), and Tomz and Van Houweling (2016)

[11]Bansak et al. (2018) does not report AMCE(Female, Male) in their paper as it is not a variable of interest for their study. We were unable to recover the replication procedure of Schwarz and Coppock (2022) and their estimates.

[12]These are Harris, Kao and Lust (2020), Horne (2020), and Schuler (2020)

[13]UC Merced experiment was removed from Crowder-Meyer et al. (2020)

(2022). The second experiment in Arnesen, Duell and Johannesson (2019) does not ask for respondents' support or vote for a candidate[14]. Therefore, we excluded this study from our sensitivity analysis, but kept it in the meta-analysis to be consistent with Schwarz and Coppock (2022).

The five additional experiments in our analysis are due to a different procedure we followed when replicating Goyal (2025), Henderson et al. (2022), Kirkland and Coppock (2018), and Crowder-Meyer et al. (2020). Schwarz and Coppock (2022) pool the datasets of separate experiments with different design features. When replicating Goyal (2025) and Kirkland and Coppock (2018), they pool the datasets of experiments with and without the party attribute. When replicating Henderson et al. (2022), they combined the sample of Republicans with Democrats, while endorsement and issue attributes differ in the tasks shown to each group. We added these experiments separately in our analysis. On the other hand, Crowder-Meyer et al. (2020) contains two experiments with MTurk respondents, and we included both. On the whole, we were able to conduct our respondent-level sensitivity analysis with 54 experiments.

We were unable to analyze contest sensitivities of 12 out of 54 experiments due to data or design-related issues. For four experiments[15], we could not find a variable that indicate questions answered by a respondent in the long form of conjoint data. Therefore, we were unable to match candidate profiles raced against each other and construct unique contest ids. The other eight experiments are not strictly forced-choice conjoint designs[16]. Thus, their replication data do not conform to the assumptions of our contest-level analysis. For example, they can feature contests among more than 2 candidates or have control profiles. Estimating the random-effects meta-analysis effects with 42 experiments, we find results close to those in Schwarz and Coppock (2022).

We also repeat the analysis with all experiments for which calculating the number of contests is possible. Whenever an experiment follows a standard conjoint design, and we do not have the replication data or contest variable to conduct the contest analysis, we calculate the expected number of unique contests as described in Section A2. As in the case of respondent analysis, if the replication data is not available, the AMCE is taken from Schwarz and Coppock (2022). For the eight experiments that are not conjoint, we computed the number of unique contests using their replication data. Estimates from these additional 17 experiments are held constant as we analyze the sensitivity of the meta-analytic effect.

[14] The experiment asks the following question: "Below, we have made up two imaginary persons/politicians of different background. We wish to know which of these you think would be mostly in agreement with the statement below." (Arnesen, Duell and Johannesson, 2019, p.45 in Appendix)

[15] Experiments in Hainmueller, Hopkins and Yamamoto (2014), and Henderson et al. (2022) and one of the experiments in Crowder-Meyer et al. (2020).

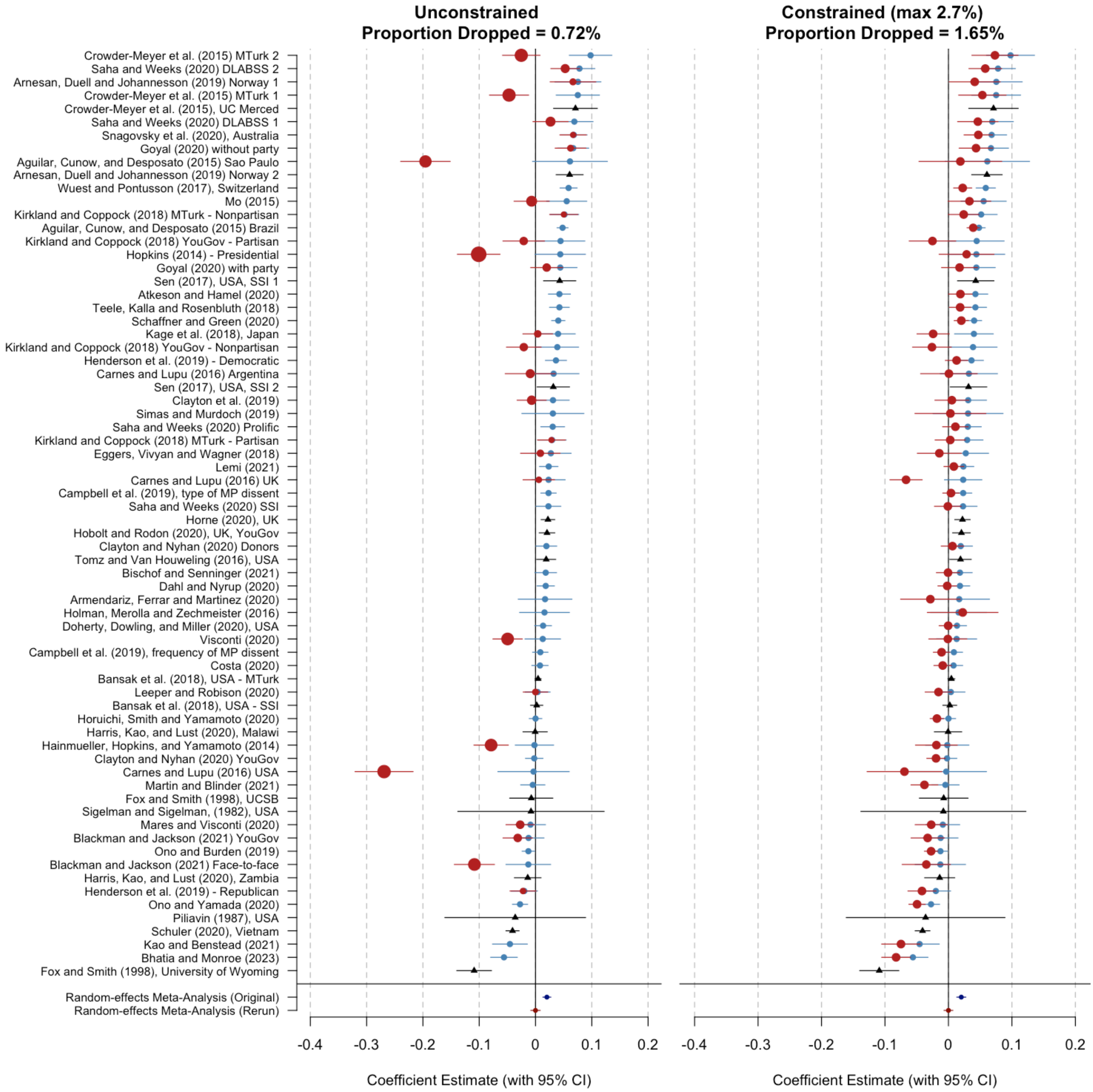
[16] Two experiments in Aguilar, Cunow and Desposato (2015), Armendariz, Farrer and Martinez (2020), Clayton et al. (2019), Holman, Merolla and Zechmeister (2016), Kao and Benstead (2021), Mo (2015), and Simas and Murdoch (2019)

Figure A1 shows the original AMCE(Female, Male) estimates in each experiment and the rerun AMCE(Female, Male) estimates after dropping the respondents to reverse the sign of the meta-analytic effect. AMCEs estimated in experiments that we can replicate and analyze are drawn in blue circles. We use black triangles for AMCEs that are kept constant. In the unconstrained case (left), it was enough to remove 0.72% of all the respondents to achieve the perturbation that reverses the sign of the meta-analysis effect. Since this procedure perturbs a few experiments disproportionately (i.e., Carnes and Lupu (2016) USA and Aguilar, Cunow and Desposato (2015) Sao Paulo), we constrain the maximum percentage of respondents that can be dropped from each experiment at 2.7% (right). This procedure results in a relatively even perturbation of each experimental sample and reverses the sign of the meta-analysis effect by removing 1.65% of all respondents.

Figure A2 shows the same procedure for the contest sensitivity of AMCE(Female, Male) using 59 experiments. Similarly, we use black triangles for AMCEs if we cannot analyze them for contest sensitivity, yet we can calculate the expected number of unique contests observed in the corresponding experiments. In the unconstrained exercises (left), removing half a percent of the contests is sufficient to reverse the sign of the meta-analysis coefficient. The smallest constraint we could find for the percentage of contests that can be dropped from each experiment is 2.3%. Under the 2.3% constraint, removing an overall 1.71% of contests is enough to reverse the sign of the meta-analysis coefficient.

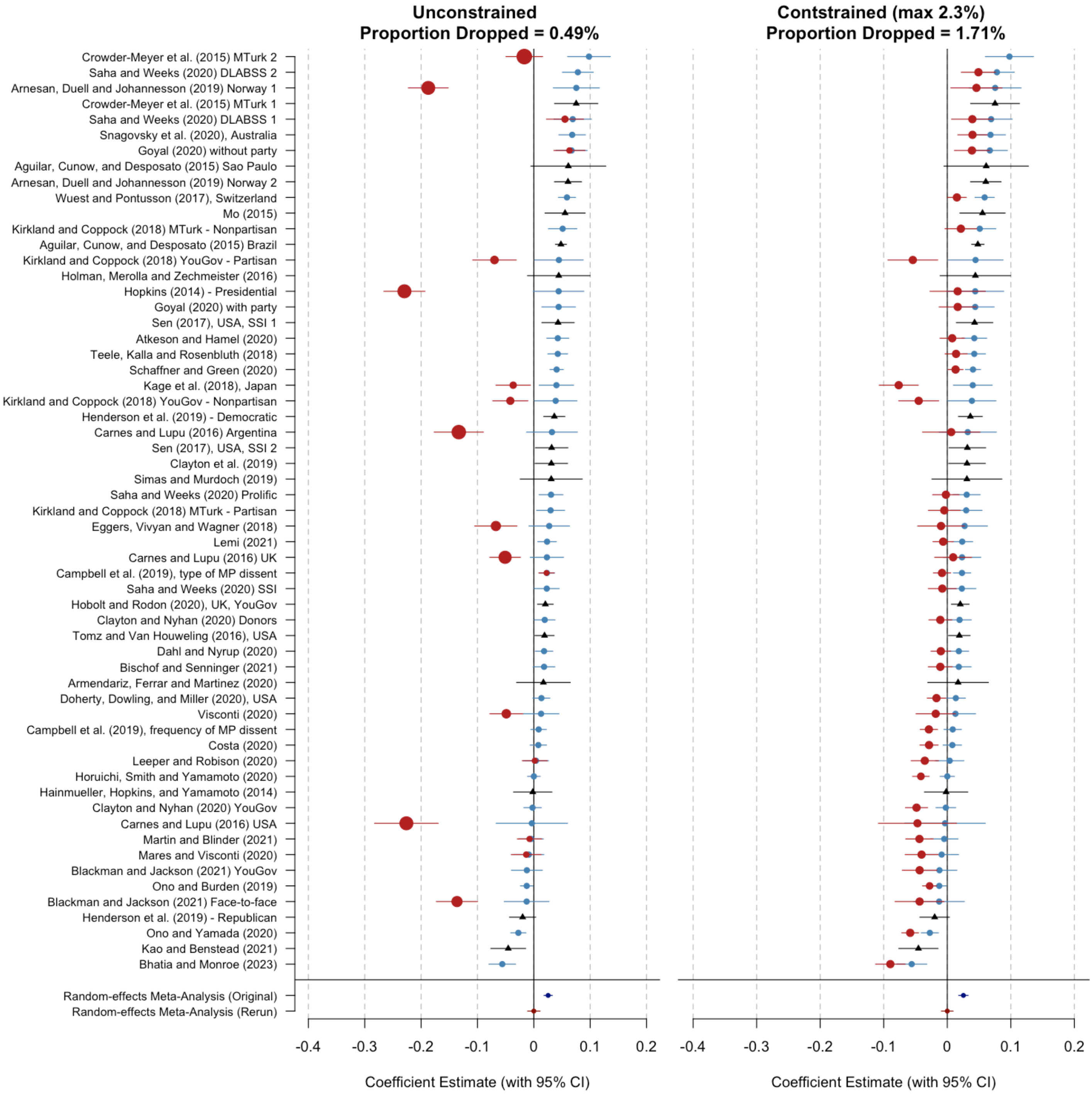
Section A5 repeats our respondent and contest analyses, excluding the experiments that we had to keep constant.

Figure A1: Respondent Analysis Targeting the Meta-analysis of AMCE(Female, Male) - with 70 Experiments



Note: Coefficient plots show how AMCE(Female, Male) changes in each experiment as we drop respondents to reverse the sign of the random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases. Coefficients from experiments that we could replicate are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate with the proportion of respondents dropped. By contrast, we take the estimates shown in black triangles from Schwarz and Coppock (2022), and they are kept constant in our analysis as we are unable to replicate those studies.

Figure A2: Contest Analysis Targeting the Meta-analysis of AMCE(Female, Male) - with 59 Experiments



Note: Coefficient plots show how AMCE(Female, Male) changes in each experiment as we drop contests to reverse the sign of the random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases. Coefficients from experiments that we could replicate are blue circles. Red circles indicate the rerun coefficients and standard errors, if any contest is dropped from those studies, where their sizes correlate with the number of contests dropped. By contrast, we take the estimates shown in black triangles from Schwarz and Coppock (2022) and they are kept constant in our analysis as we are not able to replicate those studies.

Meta-analysis of AMCE(Oldest, Youngest)

Table A2: Replication of Random-Effects Meta-analysis of AMCE(Oldest, Youngest)

	Estimate	Std.error	Conf.low	Conf.high	# Experiments
Eshima and Smith (2022)	-0.0951	0.0246	-0.143	-0.0470	16
Respondent analysis					
All (Analyzed) Experiments	-0.0877	0.0219	-0.131	-0.0448	18
Contest analysis					
All Experiments	-0.0877	0.0219	-0.131	-0.0448	18
Analyzed Experiments	-0.0843	0.0227	-0.129	-0.0398	17

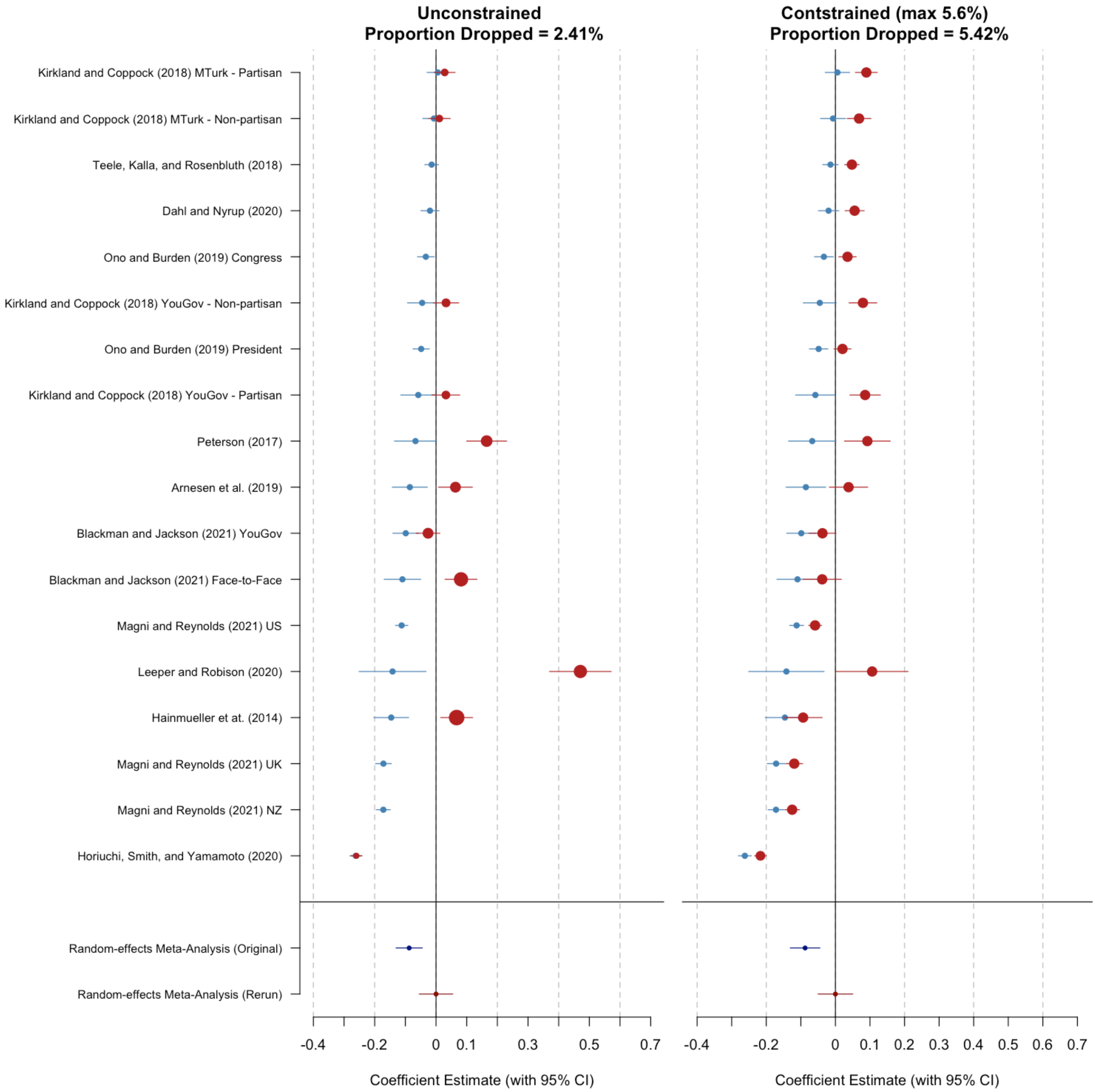
Note: Table shows our replications of the meta-analysis of AMCE(Oldest, Youngest) using the experiments included in respondent (Row 2) and contest (Row 3 and 4) analysis along with the original meta-analysis estimates in Eshima and Smith (2022) (Row 1).

Table A2 shows the results of our replication of the random-effects meta-analysis in Eshima and Smith (2022) for AMCE(Oldest, Youngest). We are able to replicate and analyze all experiments in this study. As explained above, we estimate YouGov and Mturk experiments in Kirkland and Coppock (2018) separately for partisan and non-partisan elections and therefore have 18 experiments. Similarly, since the replication data of Hainmueller, Hopkins and Yamamoto (2014) do not contain a variable indicating pairs of profiles in each contest in the experiment, we could not analyze its contest sensitivity.

Figure A3 and Figure A4 show how AMCE(Oldest, Youngest) from individual experiments change in the respondent and contest analysis as we target the sign of meta-analysis effect. Compared to the meta-analysis of AMCE(Female, Male), we need to achieve a larger perturbation to reverse the sign of the meta-analysis coefficient of AMCE(Oldest, Youngest) as its absolute value is greater (0.088). It is sufficient to remove 2.41% of the respondents in 18 experiments to achieve a perturbation of at least $|-0.088|$ (left panel of A3). This unconstrained procedure results in the removal of relatively high percentages of respondents from Hainmueller, Hopkins and Yamamoto (2014) and Blackman and Jackson (2021). The smallest constraint we could set on each experiment is 5.6%, and we can reverse the sign of the meta-analysis coefficient by removing 5.4% of all respondents and perturbing individual experiments relatively equally.

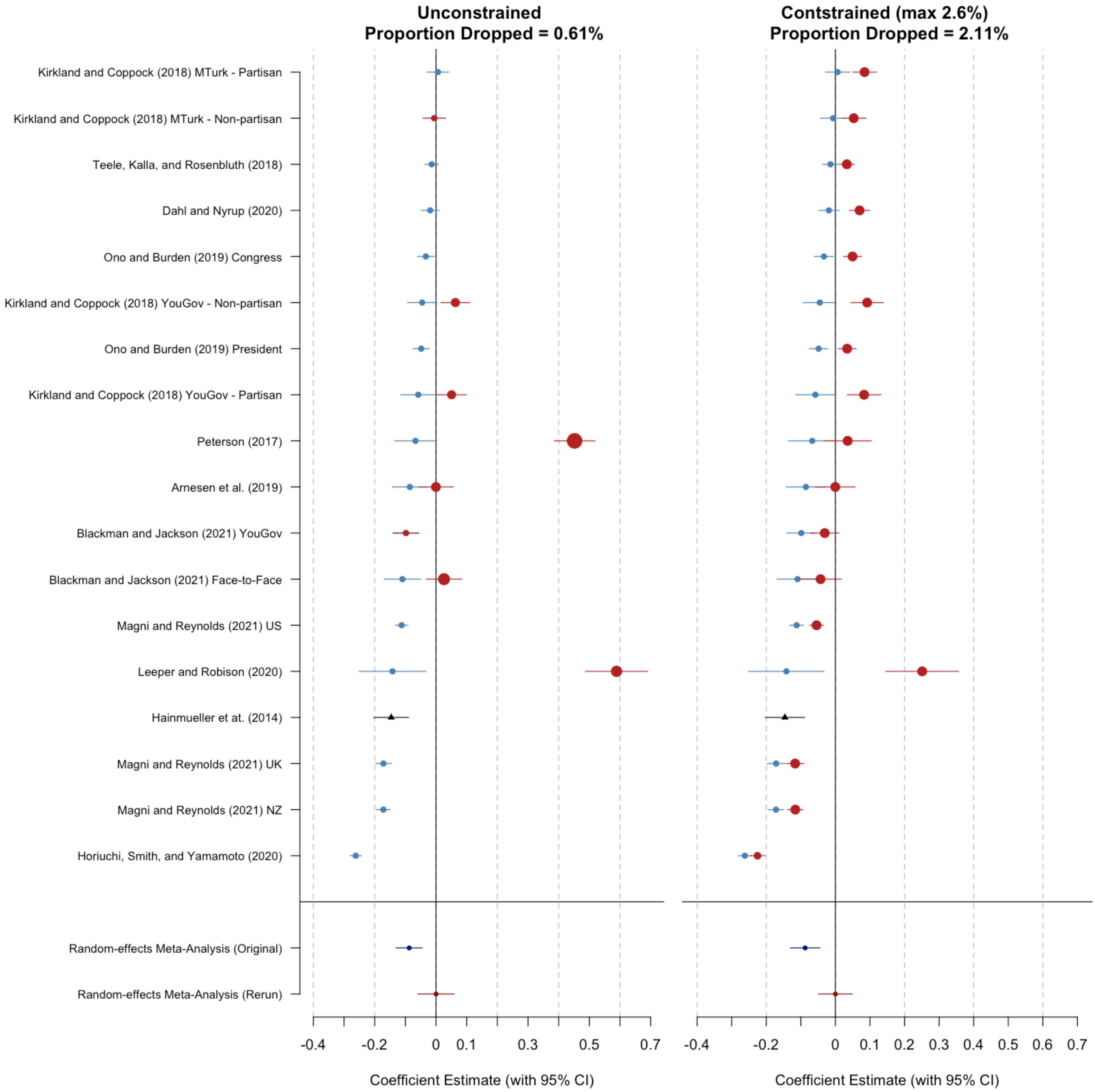
It is sufficient to remove less than 1% of the experimental contests to achieve the same perturbation and reverse the sign of the meta-analysis coefficient (Figure A4). However, to a large extent, this is done by perturbing a single experiment (Peterson (2017)). The right panel shows the results with a cap on the maximum percentage of contests that can be removed from each experiment. The minimum percentage we could find is 2.6%, and we need to remove an additional 1.5% of contests to reverse the sign of the meta-analytic effect under the constraint.

Figure A3: Respondent Analysis Targeting the Meta-analysis of AMCE(Oldest, Youngest)



Note: Coefficient plots show how AMCE(Oldest, Youngest) changes in each experiment as we drop respondents to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate the proportion of respondents dropped.

Figure A4: Contest Analysis Targeting the Meta-analysis of AMCE(Oldest, Youngest)



Note: Coefficient plots show how AMCE(Oldest, Youngest) changes in each experiment as we drop contests to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any contest is dropped from those studies, where their sizes correlate the proportion of contests dropped.

A4 Individual Experiment Results When Targeting AMCEs

This section presents the results of our sensitivity analysis conducted with individual experiments in both Schwarz and Coppock (2022) and Eshima and Smith (2022). We analyze each experiment separately by targeting the sign of AMCE estimated in the respective study.[17] Overall, most experiments appear to be highly sensitive to the removal of a small proportion of influential experiment subjects and contests. Few studies with a smaller experimental design and a large absolute effect size seem to be relatively robust. We report the results with individual experiments in this section and explore the determinants of sensitivity at the experiment and meta-analysis levels in Section A6 and the characteristics of influential respondents in Section A7.

Table A3 shows the results for the 54 replicated studies in the gender meta-analysis. The second column presents AMCE(Female, Male) in each experiment as we replicated it. The third column shows the total number of respondents in the experiment and, in parentheses, the minimum percentage of those sufficient to remove to reverse the sign of the coefficient. Individual experiments in the gender meta-analysis are highly sensitive to small changes in respondents. For half of these experiments, removing less than 3% of its respondents is sufficient to estimate AMCE(Female, Male) with the opposite sign. Only 5 experiments require a removal of 10% or more of its respondents for the reversal of the direction of AMCE(Female, Male).

The fourth column of Table A3 shows the total number of unique contests observed in each experiment; in parentheses, the percentage of contests sufficient to drop to reverse the sign of the coefficient for 42 of 54 experiments for which contest analysis is possible. The contest-level sensitivity is more pronounced. For half of the analyzed experiments, removing 1.6% or less of its unique contests is enough to achieve a perturbation sufficient to reverse the sign of AMCE(Female, Male). We need to drop more than 10% of its contests to reverse the sign of AMCE(Female, Male) in only one study, which is the second MTurk experiment in Crowder-Meyer et al. (2020). This study also has the smallest experimental dimensionality with only 36 unique contests and we need to drop 9 unique contests to reverse the sign of the effect. The same study also appeared relatively robust in our respondent-level sensitivity analysis, as it requires the removal of 12.8% of its respondents to reverse the sign of AMCE(Female, Male).

Table A4 shows the same for AMCE(Oldest, Youngest) for each experiment in Eshima and Smith (2022). Similar to the experiments analyzed for the sensitivity of AMCE(Female, Male), removing less than

[17] We observed unstable results with the R package, *zaminfluence*, provided by Giordano, Meager and Broderick (2026*a*,*b*) when estimating the influence scores for Leeper and Robison (2020) despite setting a seed number. This issue can propagate to our meta-analysis level results as the study is included in both meta-analyses. To avoid replication issues, we save the instance of sentiment scores generating the results in the paper.

3% of its respondents is sufficient to reverse the sign of AMCE(Oldest, Youngest) in 9 of 18 experiments analyzed in Eshima and Smith (2022). On the other hand, we observe a higher proportion of experiments that appear to be robust as the effect sizes are larger in these experiments. We needed to drop more than 10% of the respondents to reverse the sign of AMCE(Oldest, Youngest) in approximately one quarter of the experiments. Among the 18 estimates, AMCE(Oldest, Youngest) in Horiuchi, Smith and Yamamoto (2020) has the largest coefficient in absolute value and the lowest standard error, and requires the removal of more than 40% of respondents to reverse the sign of the effect.

Contest-level analysis is performed with 17 of 18 experiments in Table A4. Contest-level sensitivities of AMCE(Oldest, Youngest) show a pattern similar to those of AMCE(Female, Male). For half of the analyzed experiments, removing 1.5% or less of its unique contests is enough to achieve a perturbation sufficient to reverse the sign of AMCE(Oldest, Youngest). We need to drop more than 10% of unique contests only in Horiuchi, Smith and Yamamoto (2020) to reverse the sign of AMCE(Oldest, Youngest).

Table A3: Respondents and Contests Dropped to Reverse the Sign of AMCE(Female, Male) in Each Experiment

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
Aguilar, Cunow and Desposato (2015) Brazil	0.048 (0.005)	3,908 (0.111)	
Aguilar, Cunow and Desposato (2015) Sao Paulo	0.061 (0.034)	608 (0.038)	
Armendariz, Farrer and Martinez (2020)	0.017 (0.024)	1,541 (0.010)	
Arnesen, Duell and Johannesson (2019) Norway 1	0.075 (0.021)	1,106 (0.060)	1,106 (0.060)
Atkeson and Hamel (2020)	0.042 (0.010)	1,500 (0.053)	4,168 (0.029)
Bhatia and Monroe (2024)	−0.056 (0.012)	1,299 (0.063)	3,896 (0.041)
Senninger and Bischof (2021)	0.018 (0.010)	993 (0.025)	4,963 (0.014)

Continued on next page...

Table A3 – Continued from previous page

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
Blackman and Jackson (2021) Face-to-face	-0.013 (0.020)	358 (0.014)	1,432 (0.010)
Blackman and Jackson (2021) YouGov	-0.012 (0.014)	574 (0.016)	2,870 (0.009)
Campbell et al. (2019), frequency of MP dissent	0.009 (0.007)	1,899 (0.011)	7,366 (0.004)
Campbell et al. (2019), type of MP dissent	0.023 (0.007)	1,919 (0.033)	4,013 (0.016)
Carnes and Lupu (2016) Argentina	0.032 (0.023)	956 (0.027)	337 (0.030)
Carnes and Lupu (2016) UK	0.023 (0.015)	3,779 (0.004)	136 (0.044)
Carnes and Lupu (2016) USA	-0.003 (0.032)	678 (0.001)	383 (0.003)
Carey et al. (2022) Donors	0.019 (0.009)	570 (0.042)	5,516 (0.014)
Carey et al. (2022) YouGov	-0.002 (0.008)	954 (0.002)	9,270 (0.001)
Clayton et al. (2019)	0.031 (0.015)	598 (0.033)	
Costa (2021)	0.008 (0.007)	1,501 (0.012)	4,304 (0.004)
Crowder-Meyer et al. (2020) MTurk 1	0.075 (0.020)	481 (0.102)	
Crowder-Meyer et al. (2020) MTurk 2	0.098 (0.019)	494 (0.128)	36 (0.250)
Dahl and Nyrup (2021)	0.018 (0.008)	1,621 (0.024)	7,905 (0.015)
Doherty, Dowling and Miller (2019), USA	0.013 (0.008)	853 (0.027)	8,407 (0.010)
Eggers, Vivyan and Wagner (2018)	0.027 (0.018)	1,367 (0.016)	1,435 (0.016)
Goyal (2025) with party	0.044	802	2,343

Continued on next page...

Table A3 – Continued from previous page

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
	(0.015)	(0.049)	(0.036)
Goyal (2025) without party	0.067 (0.014)	842 (0.088)	2,419 (0.056)
Hainmueller, Hopkins and Yamamoto (2014)	−0.002 (0.017)	311 (0.003)	
Henderson et al. (2022) - Democratic	0.036 (0.010)	1,415 (0.045)	
Henderson et al. (2022) - Republican	−0.020 (0.012)	856 (0.025)	
Holman, Merolla and Zechmeister (2016)	0.016 (0.023)	1,001 (0.016)	
Hopkins (2014) - Presidential	0.044 (0.023)	180 (0.072)	1,194 (0.037)
Horiuchi, Smith and Yamamoto (2020)	0.00008 (0.006)	2,200 (0.0005)	22,000 (0.00005)
Kage, Rosenbluth and Tanaka (2018), Japan	0.040 (0.016)	1,611 (0.013)	4,833 (0.004)
Kao and Benstead (2021)	−0.045 (0.016)	1,490 (0.045)	
Kirkland and Coppock (2018) MTurk - Nonpartisan	0.052 (0.013)	1,164 (0.055)	2,939 (0.041)
Kirkland and Coppock (2018) MTurk - Partisan	0.030 (0.013)	1,178 (0.030)	5,984 (0.020)
Kirkland and Coppock (2018) YouGov - Nonpartisan	0.039 (0.019)	1,146 (0.012)	2,887 (0.008)
Kirkland and Coppock (2018) YouGov - Partisan	0.045 (0.022)	1,136 (0.012)	2,822 (0.008)
Leeper and Robison (2020)	0.004 (0.011)	725 (0.004)	2,810 (0.002)
Lemi (2021)	0.023 (0.008)	786 (0.047)	7,828 (0.018)
Visconti and Mares (2020)	−0.009 (0.014)	502 (0.012)	2,433 (0.006)

Continued on next page...

Table A3 – Continued from previous page

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
Martin and Blinder (2021)	-0.005 (0.011)	$3,943$ (0.004)	$3,252$ (0.002)
Mo (2015)	0.055 (0.018)	407 (0.071)	
Ono and Burden (2019)	-0.013 (0.006)	$1,583$ (0.024)	$15,830$ (0.010)
Ono and Yamada (2020)	-0.027 (0.007)	$2,686$ (0.034)	$10,654$ (0.021)
Saha and Weeks (2022) DLABSS 1	0.069 (0.017)	551 (0.091)	$1,625$ (0.057)
Saha and Weeks (2022) DLABSS 2	0.078 (0.014)	497 (0.135)	$2,447$ (0.063)
Saha and Weeks (2022) Prolific	0.031 (0.011)	869 (0.045)	$4,179$ (0.021)
Saha and Weeks (2022) SSI	0.023 (0.011)	$1,247$ (0.026)	$3,605$ (0.016)
Schaffner and Green (2019)	0.040 (0.006)	$2,936$ (0.060)	$14,442$ (0.034)
Simas and Murdoch (2019)	0.031 (0.028)	$1,263$ (0.030)	
Snagovsky et al. (2020), Australia	0.068 (0.012)	797 (0.104)	$3,462$ (0.056)
Teele, Kalla and Rosenbluth (2018)	0.042 (0.009)	$2,105$ (0.052)	$6,173$ (0.034)
Visconti (2022)	0.013 (0.016)	200 (0.025)	$1,471$ (0.009)
Wuest and Pontusson (2022), Switzerland	0.059 (0.008)	$4,520$ (0.049)	$8,849$ (0.032)

Note: Table shows the results for 54 experiments that we replicated. The second column presents our replication of $AMCE(Female, Male)$ estimated in each study. The third column shows the total number of respondents and percentage of those dropped to reverse the sign of $AMCE(Female, Male)$. The fourth column is the same for contest-level analysis. We were able to conduct our contest-level analysis for 42 experiments.

Table A4: Respondents and Contests Dropped to Reverse the Sign of AMCE(Oldest, Youngest) in Each Experiment

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
Arnesen, Duell and Johannesson (2019)	−0.085 (0.029)	1134 (0.038)	2210 (0.026)
Blackman and Jackson (2021) Face-to-Face	−0.110 (0.03)	358 (0.092)	1432 (0.046)
Blackman and Jackson (2021) YouGov	−0.099 (0.021)	574 (0.098)	2870 (0.039)
Dahl and Nyrup (2021)	−0.020 (0.015)	1621 (0.012)	7905 (0.005)
Hainmueller, Hopkins and Yamamoto (2014)	−0.146 (0.029)	311 (0.161)	
Horiuchi, Smith and Yamamoto (2020)	−0.262 (0.01)	2200 (0.427)	22000 (0.103)
Kirkland and Coppock (2018) MTurk - Non-partisan	−0.007 (0.018)	1164 (0.004)	2939 (0.003)
Kirkland and Coppock (2018) MTurk - Partisan	0.006 (0.018)	1178 (0.004)	4390 (0.002)
Kirkland and Coppock (2018) YouGov - Non-partisan	−0.045 (0.024)	1146 (0.012)	2887 (0.006)
Kirkland and Coppock (2018) YouGov - Partisan	−0.058 (0.029)	1136 (0.011)	2822 (0.006)
Leeper and Robison (2020)	−0.142 (0.056)	725 (0.026)	2810 (0.009)
Magni and Reynolds (2021) NZ	−0.172 (0.011)	1285 (0.244)	7683 (0.088)
Magni and Reynolds (2021) UK	−0.172 (0.013)	1120 (0.221)	5547 (0.09)
Magni and Reynolds (2021) US	−0.112 (0.01)	1827 (0.134)	9080 (0.054)
Ono and Burden (2019) Congress	−0.033 (0.014)	1583 (0.024)	7915 (0.01)

Continued on next page...

Table A4 – Continued from previous page

Study	Orig.Coef SE	# Respondents %Dropped	#Contests %Dropped
Ono and Burden (2019) President	-0.048 (0.014)	1583 (0.037)	7915 (0.015)
Peterson (2017)	-0.067 (0.035)	956 (0.021)	1739 (0.016)
Teele, Kalla and Rosenbluth (2018)	-0.014 (0.011)	2105 (0.011)	6173 (0.007)

Note: Table shows the results for 18 experiments that we replicated. The second column presents our replication of AMCE(Oldest, Youngest) estimated in each study. The third column shows the total number of respondents and percentage of those dropped to reverse the sign of AMCE(Oldest, Youngest). The fourth column is the same for contest-level analysis. We were able to conduct our contest-level analysis for 17 experiments.

A5 Results with Analyzed Experiments

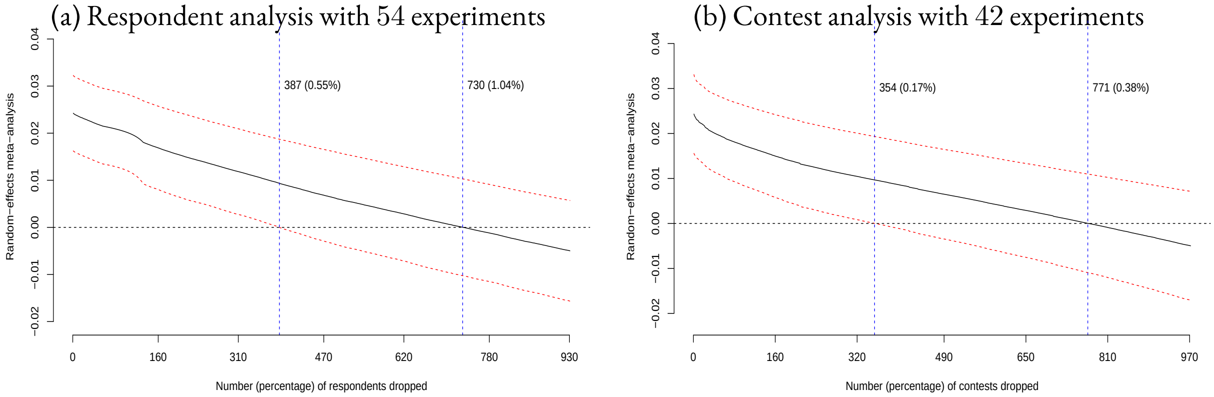
A5.1 Meta-analysis of Gender

In Section A3, we discuss that we were able to replicate 54 of the 70 experiments analyzing AMCE(Female, Male). 42 of these experiments have a standard conjoint design, and their replication packages contain variables that allow us to analyze their contest sensitivity. In this section, we repeat our main analysis of meta-analytic sensitivity excluding the experiments that we were unable to analyze individually. Our results for the meta-analysis of AMCE(Female, Male) remain stable.

Figure A5 shows the effect of dropping respondents (left) and contests (right) on meta-analysis coefficient and confidence interval. Removing a little more than 1% of all respondents (730 respondents) from 54 experiments is enough to reverse the sign of the meta-analytic effect (left panel). As presented in Figure A7 (left), this procedure can result in uneven perturbation across the studies. As we place a more restrictive constraint on the maximum proportion dropped from each experiment, we need to drop an overall higher percentage of respondents (left panel of Figure A6). The smallest constraint per experiment that allows us to achieve enough perturbation is 2.5%, and we end up dropping 2.39% of all respondents to reverse the sign of meta-analytic effect.

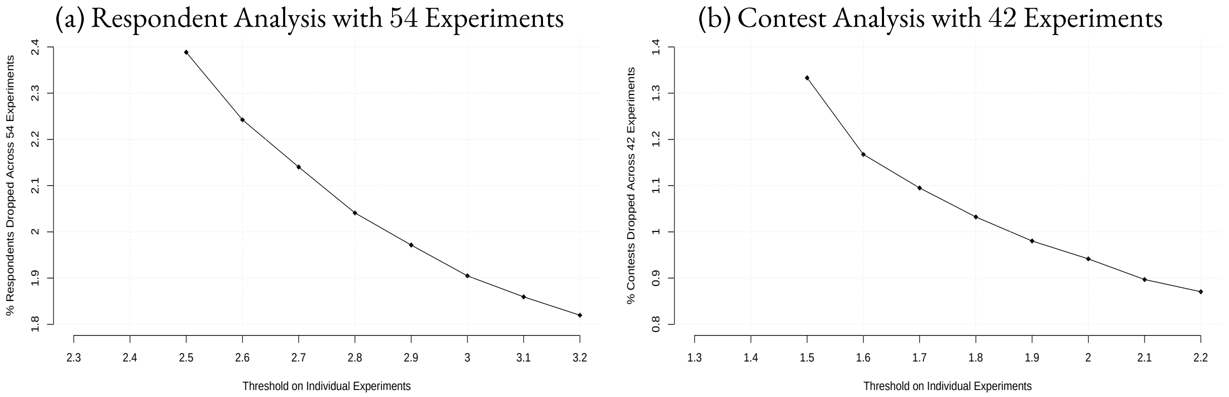
Similarly, removing 771 contests from 42 experiments is enough to reverse the sign of the meta-analysis effect, which corresponds to only 0.38% of all contests in these experiments. In the unconstrained proce-

Figure A5: Effect of Dropping Respondents and Contests on Meta-analysis Estimates



Note: Figures show how the effects change as we remove the most influential respondent (left) and contest (right) iteratively from the meta-analysis of AMCE(Female, Male) from the analyzed experiments.

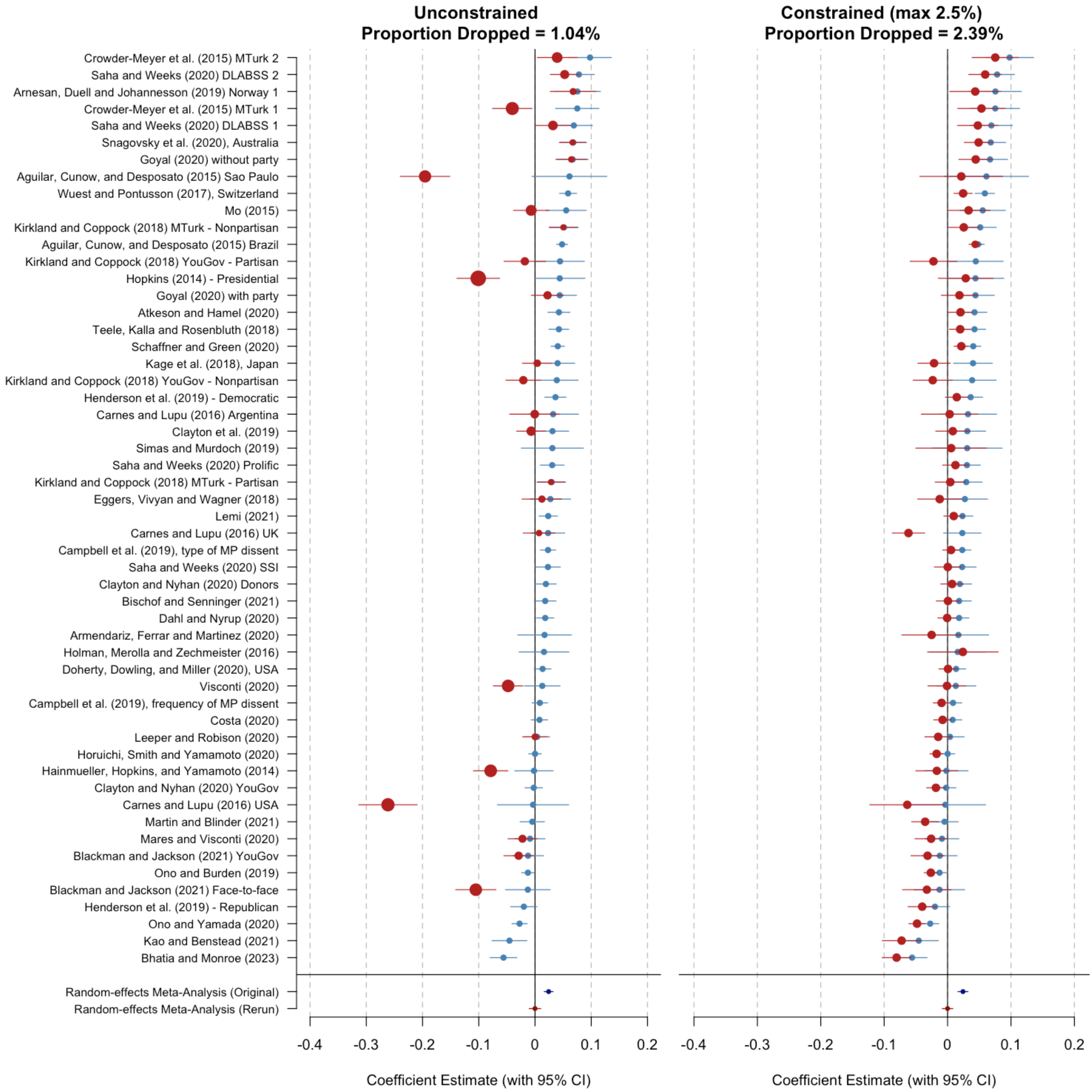
Figure A6: Change in Percentage of Respondents and Contests Dropped by Constraint



Note: Figures show how % respondents (left) and contests (right) dropped to reverse the meta-analytic effect of AMCE(Female, Male) from the analyzed experiments changes as we numerically search for the minimum cap on the maximum % for individual experiments.

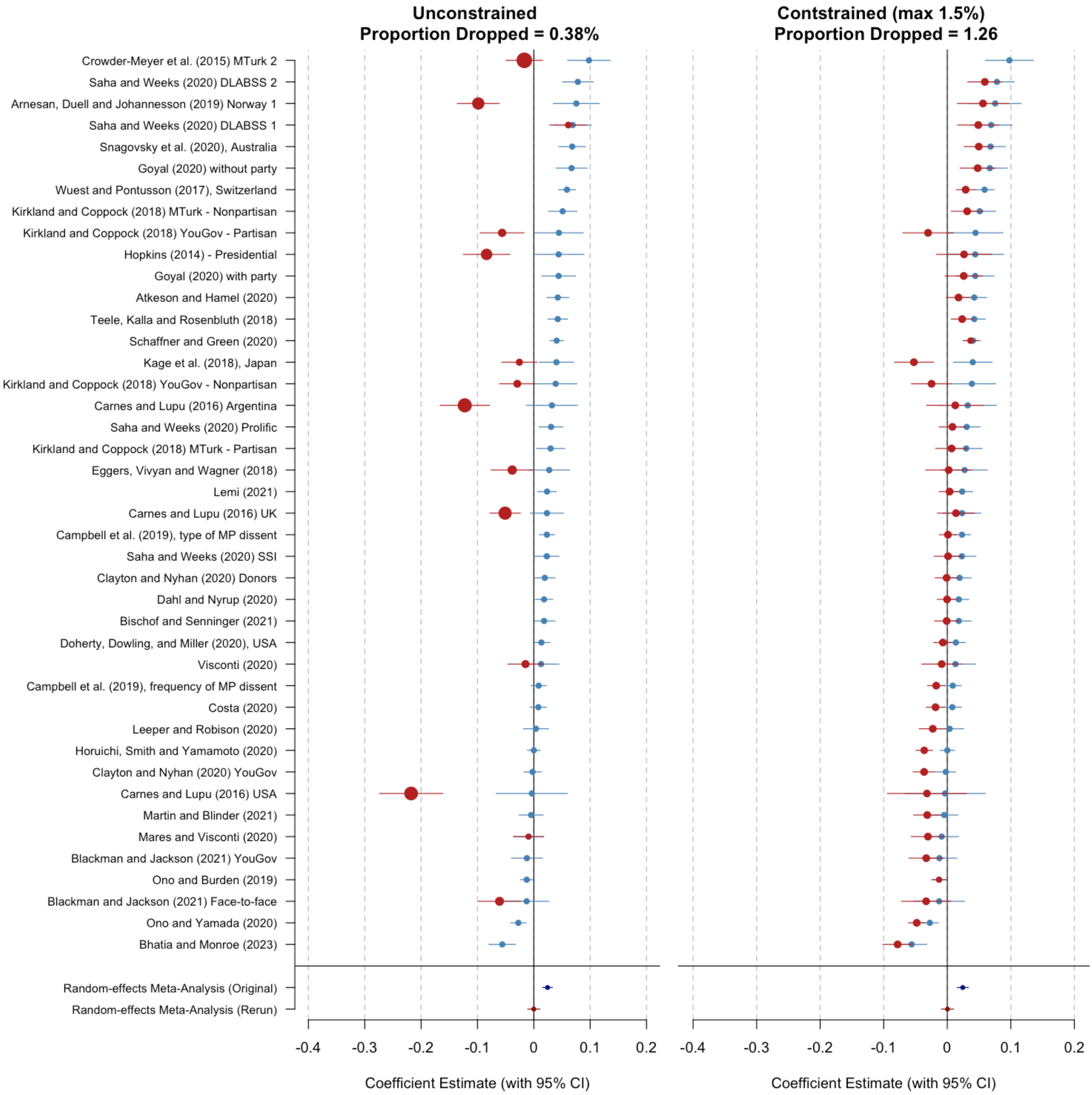
dure, we end up with removing a higher percentage of contests from the relatively low-dimensional experiments, while we drop zero contests from 28 experiments. For example, we remove 30% of contests from Crowder-Meyer et al. (2020) (MTurk2), which has only 36 unique contests. Similarly, three experiments in Carnes and Lupu (2016) have the lowest number of unique contests (less than 400 unique contests in each) after Crowder-Meyer et al. (2020) (MTurk2), and at least 17% of the contests are removed from each of these experiments. The smallest constraint on individual experiments that allows us to achieve enough perturbation to reverse the sign of the meta-analytic effect is 1.5%. At such a low threshold, we are too restrained to remove any contest from Crowder-Meyer et al. (2020) (MTurk2); however, dropping an overall 1.26% of the contests is sufficient to reverse the sign. As shown by the size of red dots in Figure A8 (right), the constrained procedure perturbs each experiment at approximately equal proportions.

Figure A7: Respondent Analysis Targeting the Meta-analysis of AMCE(Female, Male) - with 54 Experiments



Note: Coefficient plots show how AMCE(Female, Male) changes in each experiment as we drop respondents to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 54 experiments that we are able to replicate. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate the proportion of respondents dropped from each experiment.

Figure A8: Contest Analysis Targeting the Meta-analysis of AMCE(Female, Male) - with 42 Experiments



Note: Coefficient plots show how AMCE(Female, Male) changes in each experiment as we drop contests to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 42 experiments that we are able to analyze. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate the proportion of contests dropped from each experiment.

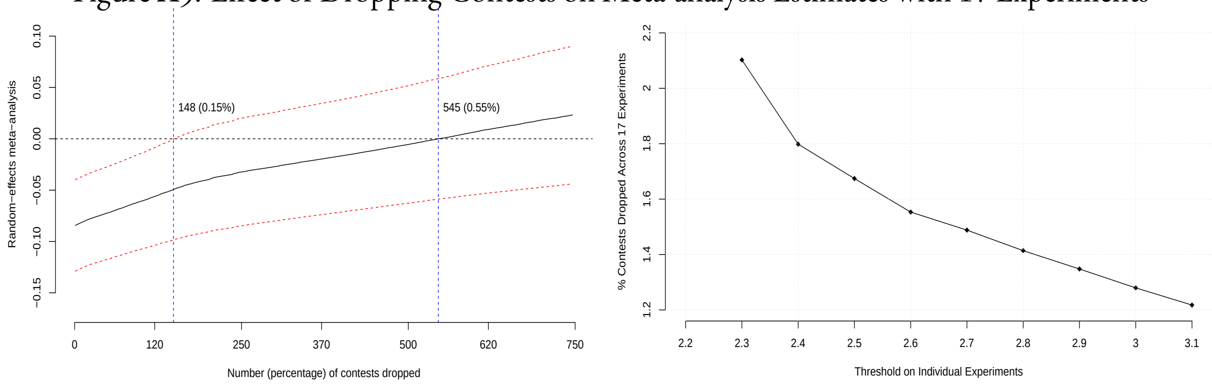
A5.2 Meta-analysis of Age

We replicated all experiments in the meta-analysis of AMCE(Oldest, Youngest); however, we were unable to conduct our contest analysis for one experiment (see Table A2). In this section, we repeat our contest-level analysis of meta-analytic sensitivity with 17 experiments excluding Hainmueller, Hopkins and Yamamoto (2014), the experiment that we were unable to analyze individually. Our results for the meta-analysis of AMCE(Oldest, Youngest) remain stable as the results are equivalent to those with 18 experiments.

The left panel of Figure A9 shows the effect of dropping unique contests iteratively on the coefficient and confidence interval of the meta-analysis. Removing 148 (0.15%) contests is sufficient to have an estimate statistically indistinguishable from zero. Once we remove about half a percent of contests, the point estimate becomes positive. The right panel explores how the proportion of contests removed to reverse the meta-analytic effect changes as we impose a threshold on the percent dropped from individual experiments. The minimum threshold under which we are still able to reverse the meta-analytic effect is 2.3%, which results in a removal of 2.1% of all contests to reverse the sign.

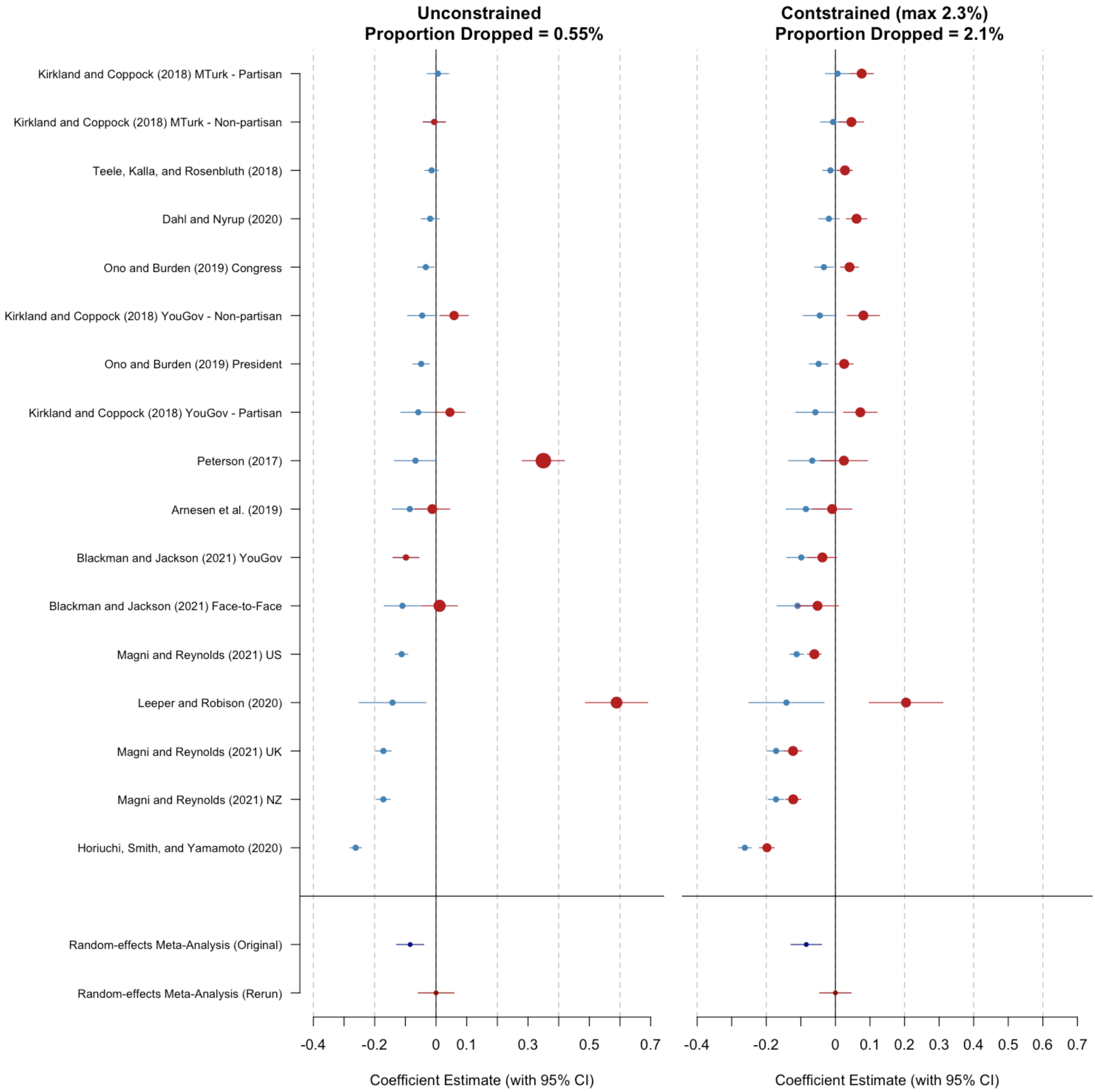
Figure A10 shows how coefficients from individual experiments are perturbed with the unconstrained (left) and constrained (right) procedures. The coefficient plots resemble those for the analysis with 18 experiments as expected since removing the experiment in Hainmueller, Hopkins and Yamamoto (2014) does not have a large impact on the magnitude of perturbation we need to achieve (as shown in Table A2, and the order of contest influences from the other experiments.

Figure A9: Effect of Dropping Contests on Meta-analysis Estimates with 17 Experiments



Note: Left figure shows how the effect changes as we remove the most influential contest iteratively from the meta-analysis of AMCE(Oldest, Youngest) using only the analyzed experiments. Right figure shows how % contests dropped to reverse the same effect changes as we numerically search for the minimum cap on the maximum % for individual experiments in the same meta-analysis.

Figure A10: Contest Analysis Targeting the Meta-analysis of AMCE(Oldest, Youngest) - with 17 Experiments



Note: Coefficient plots show how AMCE(Oldest, Youngest) changes in each experiment as we drop contests to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 17 experiments that we are able to analyze. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate the proportion of contests dropped from each experiment.

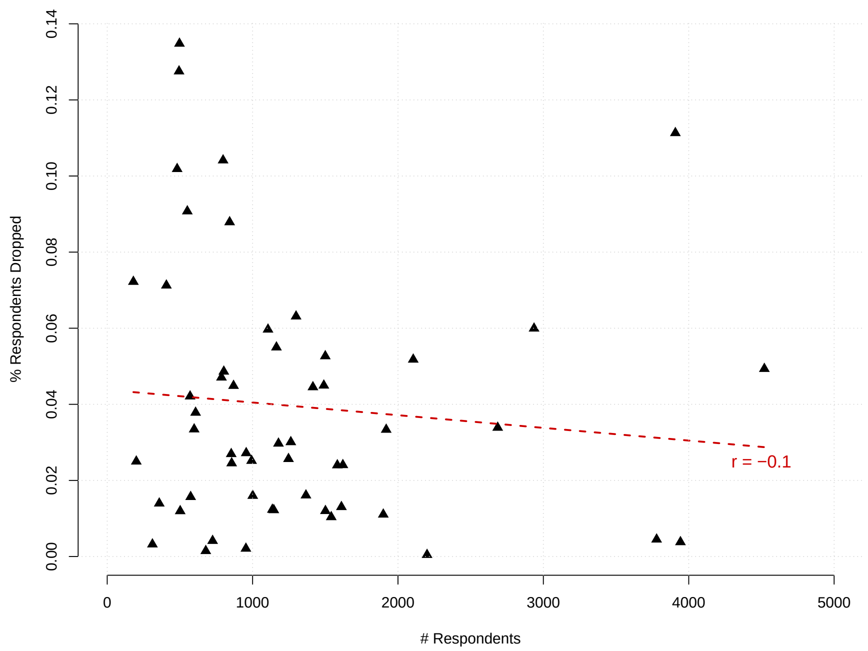
A6 What Drives the Sensitivity?

This section explores the factors associated with conjoint experiments' sensitivity to small perturbations in the sample of respondents and experimentally generated electoral contests. We focus on the meta-analysis of gender since it contains more experiments: 54 experiments analyzed for the respondent sensitivity and 42 experiments analyzed for the contest sensitivity.

Sensitivity to the Removal of Respondents

We observe that increasing the sample size does not address the issue of AMCEs' sensitivity to the removal of small fractions of respondents. Figure A11 shows that across 54 estimates of AMCE(Female, Male), there is a weakly negative and insignificant association between the number of respondents and the proportion of respondents sufficient to remove to reverse the sign of AMCE in each study. As the variation in the percentage of respondents dropped is high at any level of sample size, a visual inspection of the figure also does not indicate any clear relationship between the sample size and the minimum sufficient perturbation.

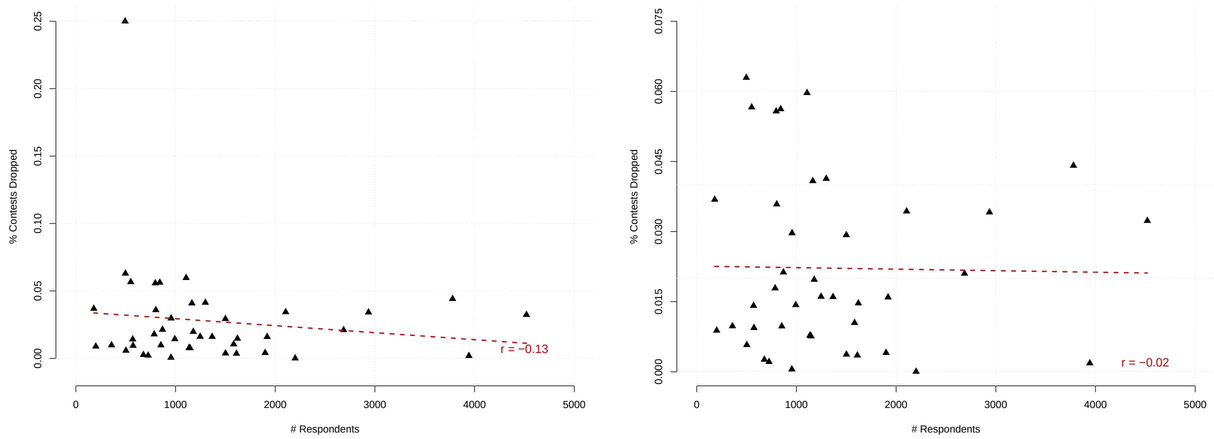
Figure A11: Percentage of Respondents Dropped to Reverse the Sign of AMCEs in Each Experiment by Number of Respondents



Note: Points on the scatter plot are 54 experiments in the gender meta-analysis, analyzed for the respondent sensitivity. The x-axis is the number of respondents in each experiment and the y-axis is the proportion of respondents dropped to reverse the sign of AMCE(Female, Male). We do not observe any association between the sample size and the sensitivity of AMCE(Female, Male).

Figure A12: Percentage of Contests Dropped to Reverse the Sign of AMCE(Female, Male) in Each Experiment by Number of Respondents

(a) 42 Experiments in Contest Analysis (b) Excluding Crowder-Meyer et al. (2020) MTurk 2



Note: Points on the scatter plots are the experiments in the gender meta-analysis, analyzed for the contest sensitivity. The x-axis is the number of respondents in each experiment and the y-axis is the proportion of contests dropped to reverse the sign of AMCE(Female, Male). Panel B shows the same, but excludes Crowder-Meyer et al. (2020) since it is an outlier on the y-axis. We do not observe any association between the sample size and the sensitivity of AMCE(Female, Male).

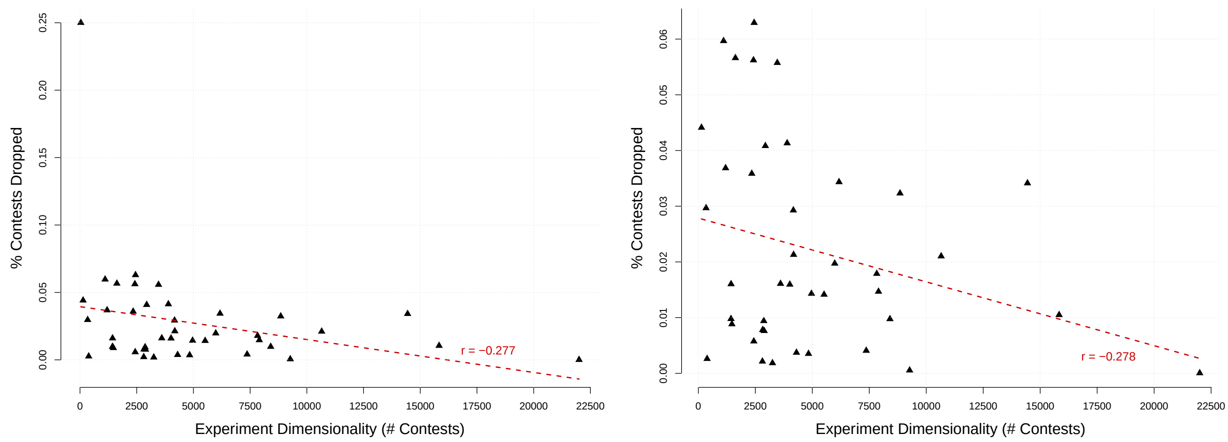
Sensitivity to the Removal of Contests

We examine whether a large sample size is related to the sensitivity of AMCEs to small perturbations in experimental contests. As in the case of respondent sensitivity, we examine the correlation between the number of respondents and the proportion of contests dropped to reverse the sign of AMCE(Female, Male) in each experiment. Figure A12 shows similar findings for the contest sensitivity of AMCEs. Across 42 analyzed experiments, we observe a weakly negative and statistically insignificant correlation between the sample size and the minimum proportion of contests that induces sufficient perturbation to reverse the sign of AMCEs. Excluding an outlier experiment, Crowder-Meyer et al. (2020), we find substantively similar results. Therefore, increasing the sample of respondents does not seem to address the sensitivity of estimands in conjoint experiments.

We then investigate how the contest sensitivity changes with respect to the observed dimensionality of conjoint tasks. We count the number of unique contests to calculate the experiment dimensionality for each experiment. Figure A13 plots the association between the dimension of experimental contests and the proportion of contests dropped to reverse the sign of AMCE(Female, Male) in each experiment. We find a negative correlation coefficient equal to -0.28. Although we do not find a strong correlation, it is significant at the 10% level with a p-value of 0.076. Excluding Crowder-Meyer et al. (2020), we similarly

Figure A13: Percentage of Contests Dropped to Reverse the Sign of AMCE(Female, Male) in Each Experiment by Experiment Dimensionality

(a) 42 Experiments in Contest Analysis (b) Excluding Crowder-Meyer et al. (2020) MTurk 2



Note: Points on the scatter plots are the experiments in the gender meta-analysis, analyzed for the contest sensitivity. The x-axis is the number of unique contests observed in each experiment and the y-axis is the proportion of those contests dropped to reverse the sign of AMCE(Female, Male). Panel B shows the same, but excludes Crowder-Meyer et al. (2020) since it is an outlier on the y-axis. We observe a negative association between the experiment dimensionality and the sensitivity of AMCE(Female, Male) significant at 10% level in both panels.

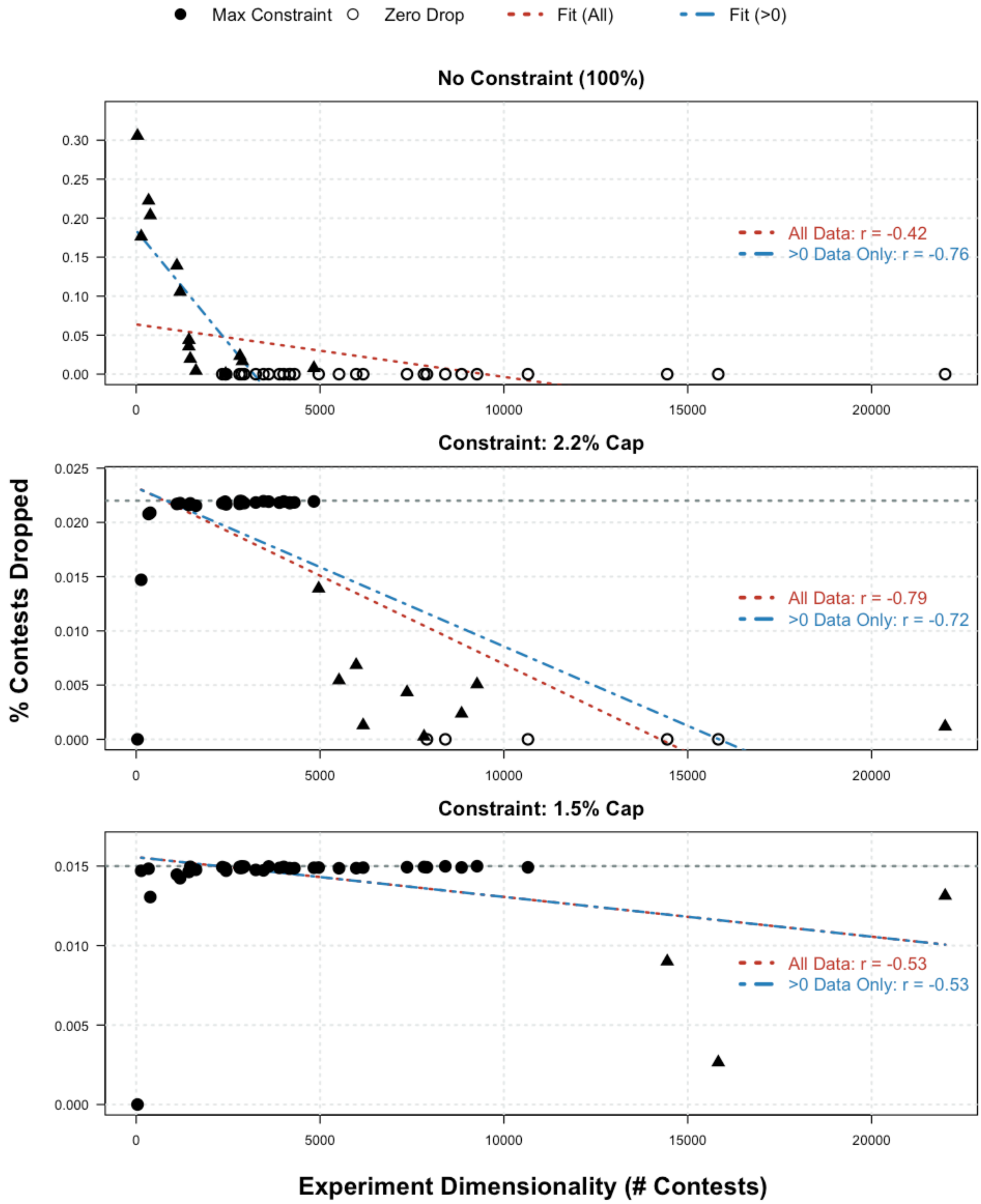
find a correlation coefficient 0.28 with a p-value of 0.078. Our analysis shows that AMCEs estimated in experiments where many candidate attributes are randomized are generally less stable to the removal of small fractions of unique contests.

We also extend our analysis to the meta-analysis level and examine which experiments the most influential contests belong to. As explained before, to find the smallest percentage of contests sufficient to reverse the sign of meta-analytical effect of being a female candidate, we remove the contests with the highest influence scores across the experiments and stop the procedure once the effect sign is reversed. We repeat our analysis with different experiment-level constraints on the percentage of unique contests that can be dropped from each experiment. Figure A14 shows that low dimensional experiments contain the most influential contests. With no cap on the experiment-level percentage (top panel), we drop contests from experiments with fewer than 5000 unique contests and the procedure stops before the high dimensional experiments are perturbed. Among the perturbed experiments, we observe a strong negative relationship between the number of unique contests and the percentage of contests dropped ($r = -0.76$). The unconstrained procedure drops 11 out of 36 contests from the experiment with the lowest dimensionality in our analysis (Crowder-Meyer et al. (2020) MTurk-2). All experiments with less than 2000 contests are perturbed by percentages strongly and negatively associated with their dimensions and overall, it is sufficient to remove 771 contests to reverse the sign of meta-analytic effect.

Constrained perturbation exercises (middle and bottom panels) corroborate our result. As we restrict the percentage of contests that can be dropped from each experiment, we observe additional experiments being perturbed following a pattern negatively correlated with the experiment dimensionality. With a 2.2% cap on each experiment, all but one experiment with less than 5000 contests reach the maximum percentage. We observe a strong negative correlation[18] (less than -0.7) similar to the unconstrained case. With a 2.2% constraint, we need to drop 1005 more contests to achieve sufficient perturbation to reverse the sign of meta-analysis coefficient. At the bottom panel, linear fit lines and Pearson correlations are less informative since nearly all experiments are censored due to the cap at 1.5%. Nevertheless, we still find a moderately strong negative correlation at -0.5 as three experiments that do not reach the maximum constraint also have the highest number of unique contests. Overall, we remove 2720 contests to reverse the meta-analytic effect.

[18]Correlation coefficients and fitted lines in the middle and bottom panels exclude Crowder-Meyer et al. (2020) MTurk-2 since the constraints at 2.2% and 1.5% do not allow us drop any contests from this experiment.

Figure A14: Percentage of Contests Dropped to Reverse the Sign of Meta-Analytic Effect of Female Candidate by Experiment Dimensionality and Constraint



Note: Figure shows the percentage of contests dropped from each experiment by their contest dimensionality when we target the sign of meta-analytic coefficient in Schwarz and Coppock (2022). At the top panel, we remove the most influential contests across 42 experiments that we can analyze their contest sensitivity. At the middle and bottom panels we impose experiment-level caps at 2.2% and 1.5% on the maximum percentage of contests that can be dropped from each experiment. In all exercises, we compute the correlation coefficient and fitted lines using all experiments (red) and only perturbed experiments (blue). Crowder-Meyer et al. (2020) MTurk 2 is excluded in the constrained cases since the 2.2% or 1.5% cap does not allow us drop any contests from this experiment, which has only 36 unique profiles. Solid (empty) circles indicate experiments whose contests are dropped at the maximum (zero) percentage.

A7 Who is Influential?

This section explores which respondents drive the conjoint results. We estimate the effect of respondent characteristics on the respondent influence scores on AMCE(Female, Male) controlling for the experiment fixed effects.

$$\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)} = \alpha_e + \alpha_1 female_{ie} + \mathbb{1}\{e \in USA\}(\alpha_2 Democrat_{ie} + \alpha_3 Republican_{ie}) + \eta_{ie} \tag{11}$$

The outcome is the influence score of respondent i, in experiment, e, standardized by the standard error of influence scores in the experiment, $\hat{\sigma}_e(\psi)$. α_e terms are the experiment fixed effects. $female_{ie}$ indicates whether a respondent is female. We were able to find data for the gender of the respondents in 45 of 54 experiments. For experiments conducted in the US, $Democrat_{ie}$ and $Republican_{ie}$ indicate whether a respondent supports or leans towards the Democratic or Republican Party, where the baseline is independent or other. We have the party of the respondents in 24 experiments.

The confidence intervals are estimated with the stratified bootstrapped estimation. At each iteration, we sample respondents from each experiment with replacement, re-estimate the influence scores targeting AMCE(Female, Male) in the corresponding experiment, and estimate the Equation 11. When sampling respondents, we use sample size equal to the number of respondents in each study.

Table A5 shows our regression results. We find that female respondents have a positive and significant influence on the AMCE of female vs. male candidates. On the other hand, being a Republican respondent as opposed to being independent is associated with a 0.103 decline in the respondent's standardized influence scores on AMCE(Female, Male). Female and Republican respondents have an effect equivalent to approximately 5% and 10% of the standard deviation in the standardized influence scores, respectively. Nevertheless, the average standardized respondent influence on AMCE(Female, Male) is near zero in all specifications, while the minimum and maximum influences are about four to fourteen standard deviations away[19]. Our analysis indicates that some respondents have intense preferences for female or male candidates with outlier influences on AMCE(Female, Male), while the median respondent has a small negative influence in each specification. We argue that since respondents can have intense preferences for some attribute levels, and outlier influences on corresponding AMCEs, it is possible that AMCEs indicate aggregate

[19] Summary statistics for the standardized influence scores in 54 analyzed experiments is similar to the subsets of experiments used in the regressions: $E(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)}) = 2.27 \times 10^{-16}$, $median(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)}) = -0.015$, $min(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)}) = -14.46$ and $max(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)}) = 12.56$

Table A5: Effect of Respondent Characteristics on Respondent Influence

Outcome:	*Standardized Respondent Influence Scores on AMCE(Female, Male)*		
	1.	2.	3.
Female	0.058^*		0.053^*
	$[0.035; 0.093]$		$[0.008; 0.108]$
Democrat		0.027	0.046
		$[-0.037; 0.095]$	$[-0.025; 0.130]$
Republican		-0.103^*	-0.104^*
		$[-0.189; -0.045]$	$[-0.200; -0.031]$
Num. obs.	58670	24114	17886
Num. Experiments	45	24	19
Sample	Full	USA	USA
Experiment FE	Yes	Yes	Yes
$E(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)})$	0.00002	-0.00010	-0.00029
$SD(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)})$	0.999	1.000	0.999
$min(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)})$	-14.463	-9.646	-5.401
$max(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)})$	11.815	12.555	4.382
$median(\frac{\hat{\psi}_{ie}}{\hat{\sigma}_e(\psi)})$	-0.017	-0.018	-0.022
R^2 (full model)	0.001	0.003	0.004
Adj. R^2 (full model)	0.000	0.002	0.003

Note: Table shows the regression results with the Equation 11. 95% bootstrap confidence intervals in parentheses. We use stratified bootstrap and sample respondents from each study with replacement. Sample size is equal to the number of respondents in each study.

preferences different from the median influence score. Accounting for the intensity of preferences makes the AMCE estimates unstable since dropping a small fraction of respondents with outlier influence scores would be sufficient to reverse the results.

A8 Results for AMCE(Oldest, SecondOldest)

We repeat our analysis for AMCE(Oldest, SecondOldest) in Eshima and Smith (2022). Table A6 shows our replication of the meta-analysis. We are able to replicate and conduct our respondent-level analysis for all experiments. However, we cannot conduct our contest-level analysis for Hainmueller, Hopkins and Yamamoto (2014) as their replication data do not contain a variable for profile pairs at each question. As in our analysis of AMCE(Oldest, Youngest), we estimate YouGov and Mturk experiments in Kirkland and Coppock (2018) separately for partisan and non-partisan elections and therefore have 2 more experiments.

Our procedure closely replicates the original meta-analysis coefficient and standard error in Eshima and Smith (2022). Relative to the meta-analytic effect of being the oldest candidate compared to the youngest

Table A6: Replication of Random-effects Meta-analysis of AMCE(Oldest, SecondOldest)

	Estimate	Std.error	Conf.low	Conf.high	# Experiments
Eshima and Smith (2022)	-0.0661	0.0159	-0.0973	-0.0349	16
Respondent analysis					
All (Analyzed) Experiments	-0.0649	0.0142	-0.0928	-0.0371	18
Contest analysis					
All Experiments	-0.0649	0.0142	-0.0928	-0.0371	18
Analyzed Experiments	-0.0639	0.0148	-0.0928	-0.0349	17

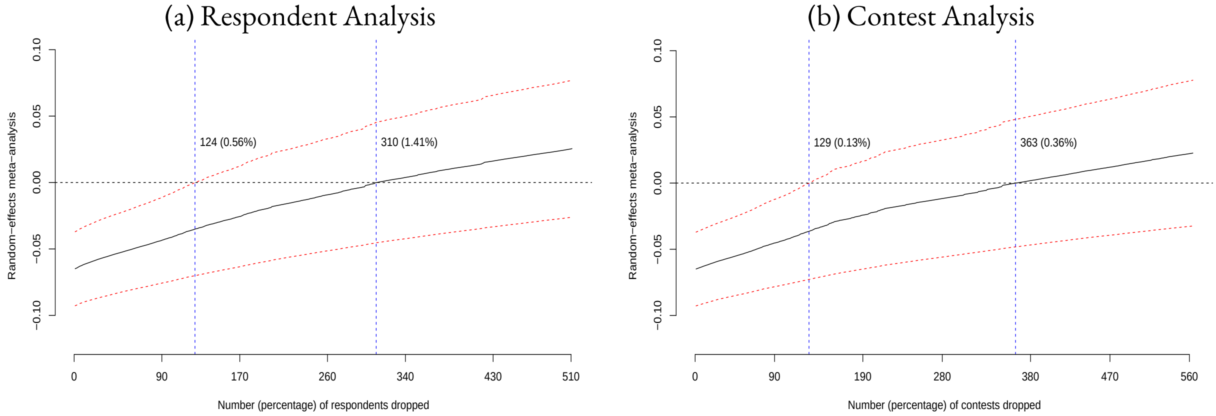
Note: Table shows our replications of the meta-analysis of AMCE(Oldest, SecondOldest) using the experiments included in respondent (Row 2) and contest (Row 3 and 4) analysis along with the original meta-analysis estimates in Eshima and Smith (2022) (Row 1).

(-0.095), the effect size is smaller with the second oldest baseline (-0.066). Overall, we find that the meta-analysis with the second oldest baseline is more sensitive than that with the youngest baseline to the removal of small fractions of respondents, while the results for the contest analysis are similar.

Figure A15 shows how the meta-analytic estimates change as we remove influential respondents (left) and contests (right) from 18 experiments when there is no constraint. Removing only 310 (1.41%) of respondents is sufficient to reverse the meta-analytic effect, while we needed to remove 531 (2.41%) with the youngest baseline. This procedure removes more respondents from Leeper and Robison (2020), Blackman and Jackson (2021), and Hainmueller, Hopkins and Yamamoto (2014), and seven studies remain unperturbed as shown in Figure A18. Nevertheless, the sensitivity of the meta-analysis estimates is not driven by a few experiments. When we constrain the percentage of respondents that can be removed from individual experiments at 4% (the minimum cap we could find numerically), dropping 3.86% of all respondents is sufficient to reverse the sign of the meta-analysis coefficient. When analyzed individually, a majority of the experiments included in the meta-analysis appear to be highly sensitive to small perturbations in the sample of respondents. By targeting AMCE(Oldest, Second Oldest) in individual experiments, we can reverse the effect in 10 of 18 experiments by removing less than 3% of their respondents (Table A7).

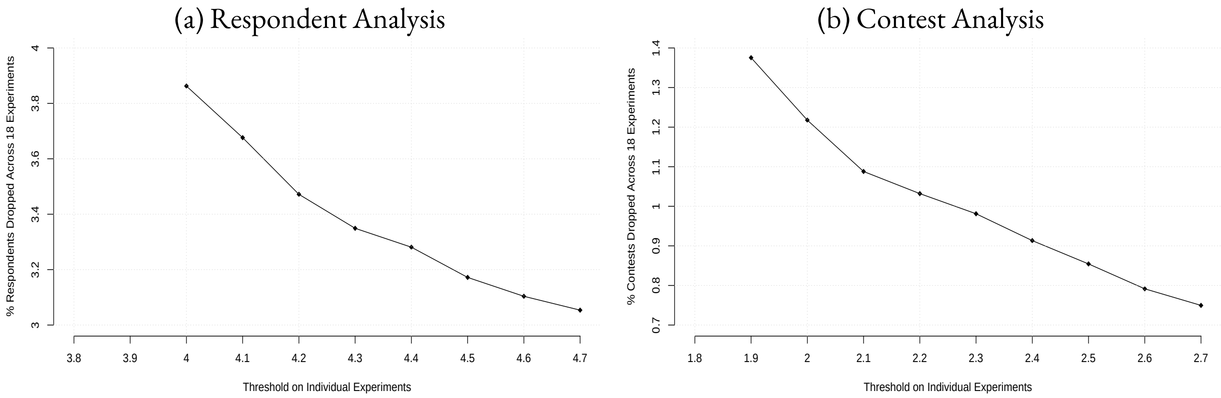
Contest-level sensitivities are even more profound. Removing 0.36% of all contests is sufficient to reverse the sign of the meta-analysis coefficient. Under a 1.9% constraint on the percentage of contests that can be removed from individual experiments, it is sufficient to drop less than 1.5% of contests to reverse the meta-analytic effect by perturbing individual experiments relatively evenly. Results remain substantively equivalent excluding Hainmueller, Hopkins and Yamamoto (2014) that we were not able to analyze individually (Figures A17 and A20).

Figure A15: Effect of Dropping Respondents and Contests on Meta-analysis Estimates of AMCE(Oldest, SecondOldest) - with 18 Experiments



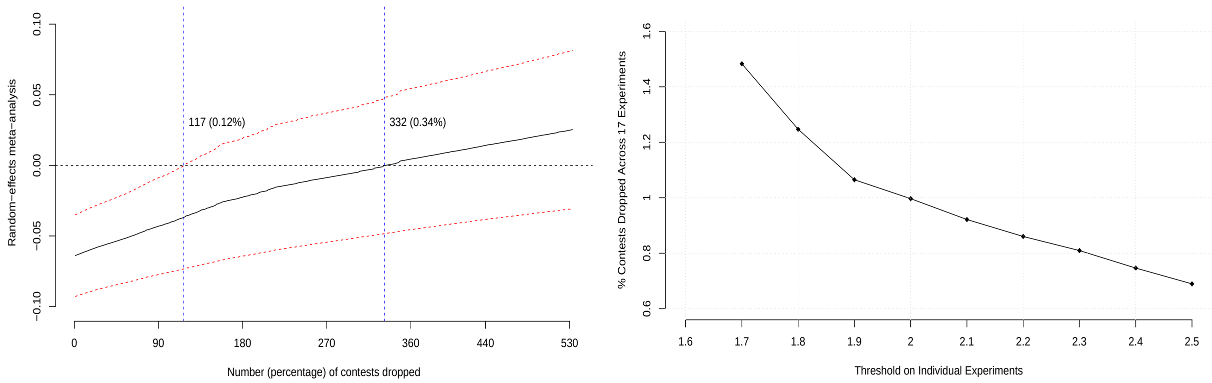
Note: Figures show how the effects change as we remove the most influential respondent (left) and contest (right) iteratively from the meta-analysis of AMCE(Oldest, SecondOldest).

Figure A16: Change in Percentage of Respondents and Contests Dropped to Reverse the Sign of Meta-analytic Effect by Constraint - with 18 Experiments



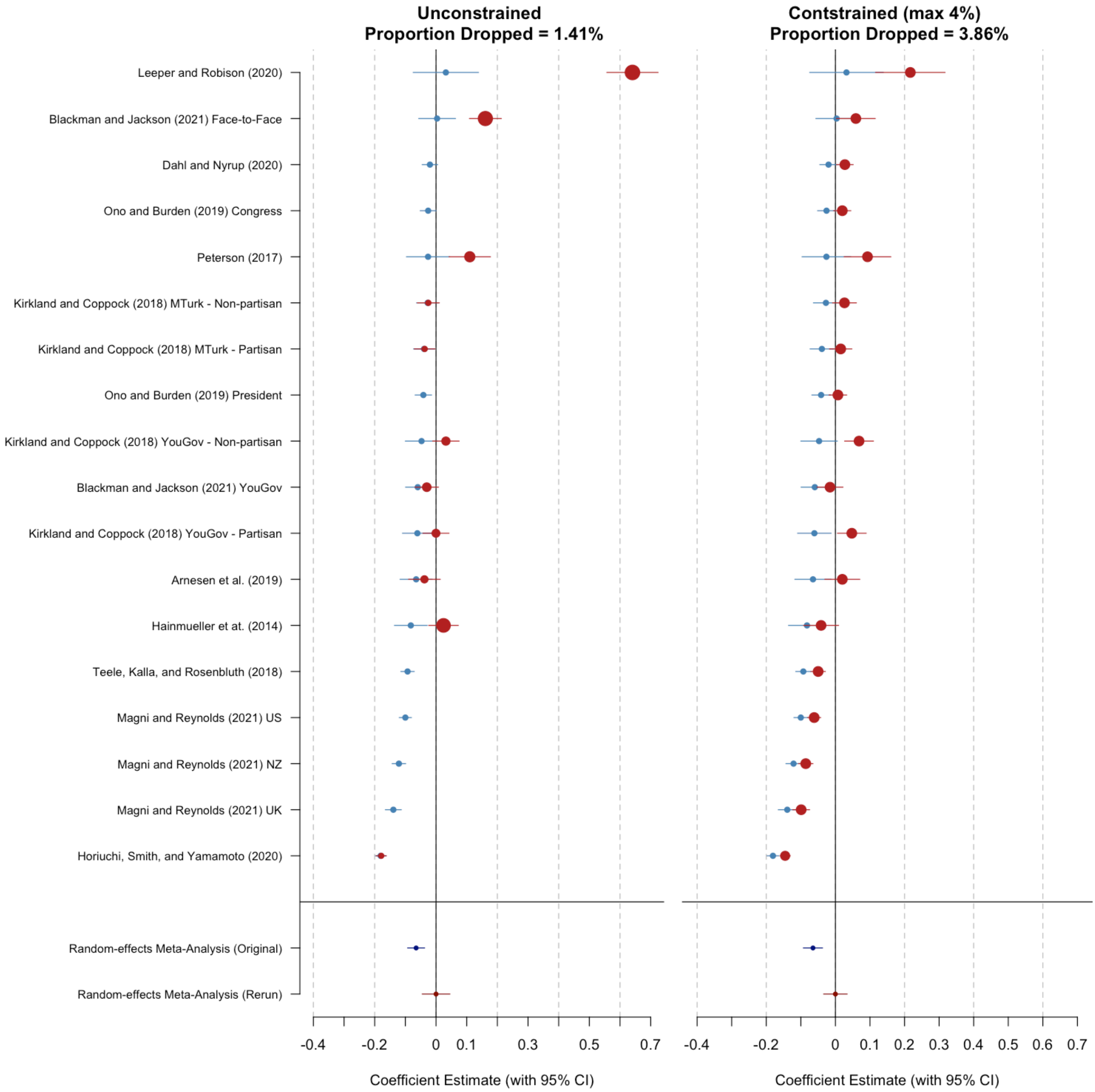
Note: Figures show how % respondents (left) and contests (right) dropped to reverse the meta-analytic effect of AMCE(Oldest, SecondOldest) changes as we numerically search for the minimum cap on the maximum percentage for individual experiments.

Figure A17: Effect of Dropping Contests on Meta-analysis Effect AMCE(Oldest, SecondOldest) - with 17 Experiments



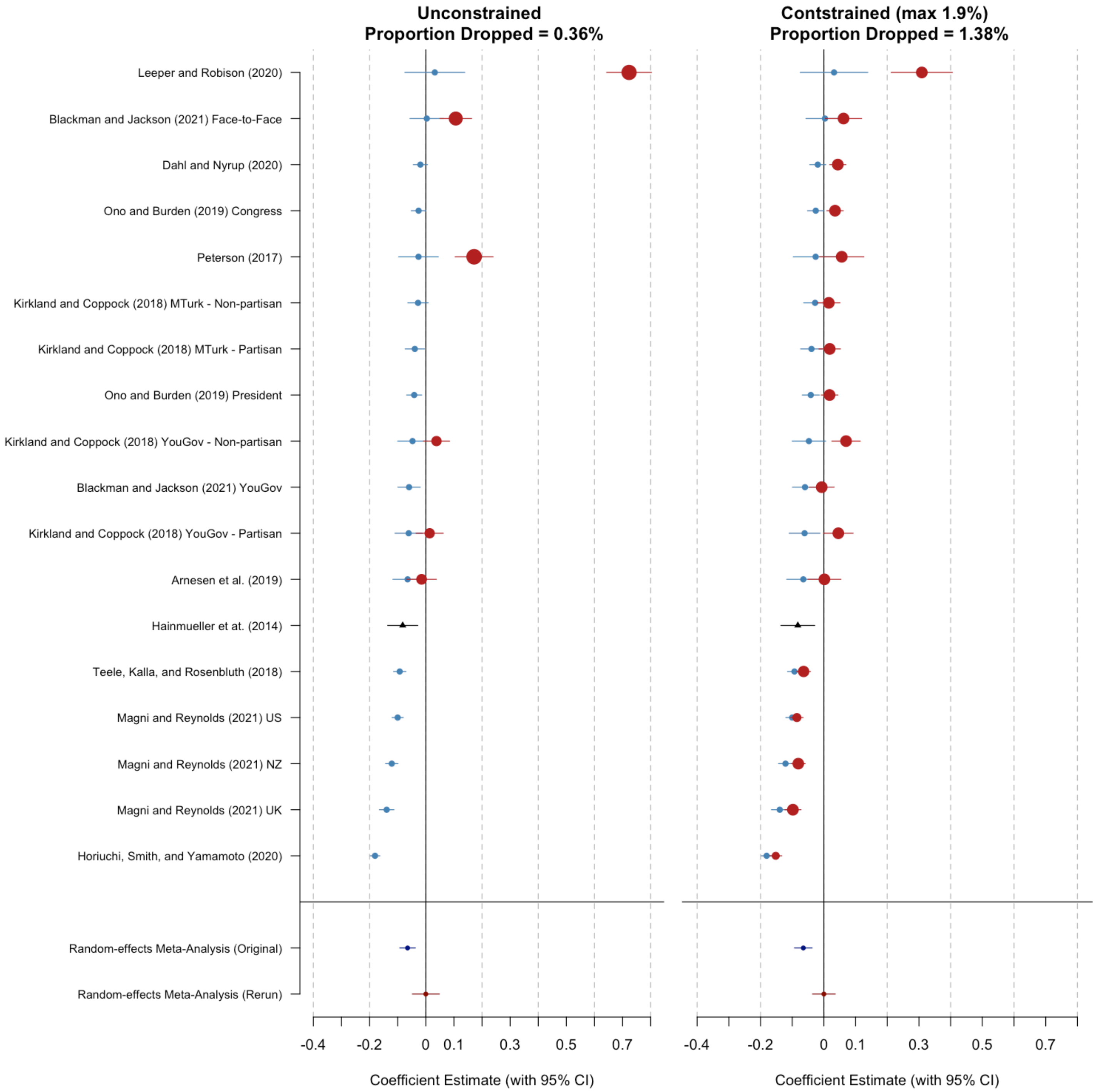
Note: Left figure shows how the effect changes as we remove the most influential contest iteratively from the meta-analysis of AMCE(Oldest, SecondOldest) using only the analyzed experiments. Right figure shows how % contests dropped to reverse the same effect changes as we numerically search for the minimum cap on the maximum % for individual experiments in the same meta-analysis.

Figure A18: Respondent Analysis Targeting the Meta-analysis of AMCE(Oldest, SecondOldest) - with 18 Experiments



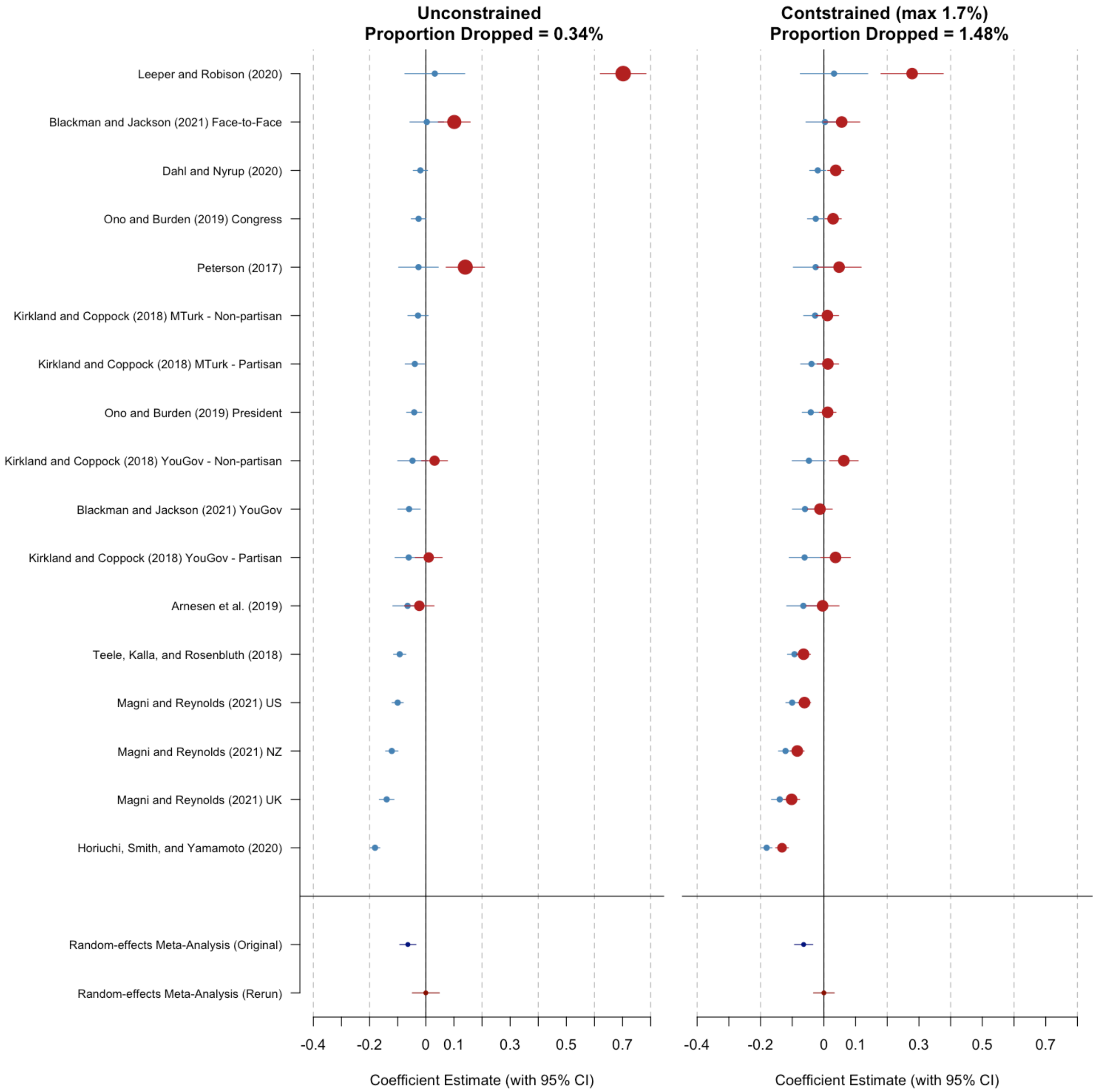
Note: Coefficient plots show how AMCE(Oldest, SecondOldest) changes in each experiment as we drop respondents to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 18 experiments. Original AMCEs are blue circles. Red circles show the rerun coefficients and standard errors if any respondent is dropped from those studies, where their sizes correlate the proportion of contests dropped from each experiment.

Figure A19: Contest Analysis Targeting the Meta-analysis of AMCE(Oldest, SecondOldest) - with 18 Experiments



Note: Coefficient plots show how AMCE(Oldest, SecondOldest) changes in each experiment as we drop contests to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 18 experiments. Original AMCEs that we are able to analyze are blue circles. Red circles show the rerun coefficients and standard errors if any contest is dropped from those studies, where their sizes correlate the proportion of contests dropped from each experiment.

Figure A20: Contest Analysis Targeting the Meta-analysis of AMCE(Oldest, SecondOldest) - with 17 Experiments



Note: Coefficient plots show how AMCE(Oldest, SecondOldest) changes in each experiment as we drop contests to reverse the sign of random-effects meta-analysis coefficient in the unconstrained (left) and constrained (right) cases using 17 experiments that we are able to analyze. Original AMCEs that we are blue circles. Red circles show the rerun coefficients and standard errors if any contest is dropped from those studies, where their sizes correlate the proportion of contests dropped from each experiment.

Table A7: Respondents and Contests Dropped to Reverse the Sign of AMCE(Oldest, SecondOldest) in Each Experiment

Study	Orig.Coef SE	# Respondents % Dropped	# Contests % Dropped
Arnesen et al. (2019)	-0.065	1134	2210
	(0.027)	(0.03)	(0.018)
Blackman and Jackson (2021) Face-to-Face	0.003	358	1432
	(0.031)	(0.003)	(0.001)
Blackman and Jackson (2021) YouGov	-0.060	574	2870
	(0.02)	(0.056)	(0.022)
Dahl and Nyrup (2020)	-0.020	1621	7905
	(0.013)	(0.015)	(0.006)
Hainmueller et at. (2014)	-0.082	311	
	(0.027)	(0.093)	
Horiuchi, Smith, and Yamamoto (2020)	-0.181	2200	22000
	(0.009)	(0.298)	(0.057)
Kirkland and Coppock (2018) MTurk - Non-partisan	-0.027	1164	2939
	(0.018)	(0.017)	(0.012)
Kirkland and Coppock (2018) MTurk - Partisan	-0.039	1178	4395
	(0.017)	(0.027)	(0.013)
Kirkland and Coppock (2018) YouGov - Non-partisan	-0.047	1146	2887
	(0.027)	(0.009)	(0.005)
Kirkland and Coppock (2018) YouGov - Partisan	-0.061	1136	2822
	(0.025)	(0.018)	(0.009)
Leeper and Robison (2020)	0.032	725	2810
	(0.054)	(0.007)	(0.002)
Magni and Reynolds (2021) NZ	-0.121	1285	7683
	(0.011)	(0.168)	(0.06)
Magni and Reynolds (2021) UK	-0.139	1120	5547
	(0.013)	(0.169)	(0.069)
Magni and Reynolds (2021) US	-0.100	1827	9080
	(0.01)	(0.122)	(0.047)
Ono and Burden (2019) Congress	-0.026	1583	7915
	(0.013)	(0.021)	(0.008)
Ono and Burden (2019) President	-0.041	1583	7915
	(0.014)	(0.033)	(0.013)
Peterson (2017)	-0.026	956	1739
	(0.036)	(0.007)	(0.006)
Teele, Kalla, and Rosenbluth (2018)	-0.093	2105	6173
	(0.011)	(0.096)	(0.056)

Note: Table shows the results for 18 experiments that we replicated. The second column presents our replication of AMCE(Oldest, SecondOldest) estimated in each study. The third column shows the total number of respondents and percentage of those dropped to reverse the sign of AMCE(Oldest, SecondOldest). The fourth column is the same for contest-level analysis. We were able to conduct our contest-level analysis for 17 experiments.

Appendix References

Aguilar, Rosario, Saul Cunow and Scott Desposato. 2015. "Choice sets, gender, and candidate choice in Brazil." *Electoral Studies* 39:230–242.

Armendariz, Karen Caballero, Ben Farrer and Monica Martinez. 2020. "Badge of Courage or Sign of Criminality? Experimental Evidence for How Voters Respond to Candidates with a Political Arrest in Their Past." *Social Science Quarterly* 101:2203–2219.

Arnesen, Sveinung, Dominik Duell and Mikael Poul Johannesson. 2019. "Do citizens make inferences from political candidate characteristics when aiming for substantive representation?" *Electoral Studies* 57:46–60.
URL: *https://www.sciencedirect.com/science/article/pii/S0261379418302257*

Atkeson, Lonna Rae and Brian T. Hamel. 2020. "Fit for the Job: Candidate Qualifications and Vote Choice in Low Information Elections." *Political Behavior* 41(1).

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2018. "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments." *Political Analysis* 26:112–119.

Bhatia, Jasmine and Steve L. Monroe. 2024. "Insecurity and Support for Female Leadership in Conflict States: Evidence from Afghanistan." *British Journal of Political Science* 54:771–792.

Blackman, Alexandra and Marlette Jackson. 2021. "Gender Stereotypes, Political Leadership, and Voting Behavior in Tunisia." *Political Behavior* 43:1037–66.

Campbell, Rosie, Philip Cowley, Nick Vivyan and Markus Wagner. 2019. "Legislator Dissent as a Valence Signal." *British Journal of Political Science* 49(1):105–28.

Carey, John, Katherine Clayton, Gretchen Helmke, Brendan Nyhan, Mitchell Sanders and Susan Stokes. 2022. "Who Will Defend Democracy? Evaluating Tradeoffs in Candidate Support Among Partisan Donors and Voters." *Journal of Elections, Public Opinion and Parties* 32(1):230–245.

Carnes, Nicholas and Noam Lupu. 2016. "Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class." *American Political Science Review* 110(4):832–844.

Clayton, Amanda, Amanda Lea Robinson, Martha C. Johnson and Ragnhild Muriaas. 2019. "(How) Do Voters Discriminate against Women Candidates? Experimental and Qualitative Evidence from Malawi." *Comparative Political Studies* 53(3-4):601–630.

Costa, Mia. 2021. "Ideology, Not Affect: What Americans Want from Political Representation." *American Journal of Political Science* 65(2):342–358.

Crowder-Meyer, Melody, Shana Kushner Gadarian, Jessica Trounstine and Kau Vue. 2020. "A Different Kind of Disadvantage: Candidate Race, Cognitive Complexity, and Voter Choice." *Political Behavior* 42:509–530.

Dahl, Malte and Jacob Nyrup. 2021. "Confident and cautious candidates: Explaining under-representation of women in Danish municipal politics." *European Journal of Political Research* 60(1):199–224.

Doherty, David, Conor M. Dowling and Michael G. Miller. 2019. "Do Local Party Chairs Think Women and Minority Candidates Can Win? Evidence from a Conjoint Experiment." *The Journal of Politics* 81(4):1282–1297.

Eggers, Andrew C., Nick Vivyan and Markus Wagner. 2018. "Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life?" *Journal of Politics* 80(1):321–326.

Eshima, Shusei and Daniel M. Smith. 2022. "Just a Number? Voter Evaluations of Age in Candidate-Choice Experiments." *The Journal of Politics* 84(3):1856–1861.

Fox, Richard L. and Eric R. A. N. Smith. 1998. "The Role of Candidate Sex in Voter Decision-Making." *Political Psychology* 19(2):405–419.

Giordano, Ryan, Rachael Meager and Tamara Broderick. 2026*a*. "An automatic finite-sample robustness metric: when can dropping a little data change conclusions? Part I: definitions and experiments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 384(2321):20250001.

Giordano, Ryan, Rachael Meager and Tamara Broderick. 2026*b*. "An automatic finite-sample robustness metric: when can dropping a little data change conclusions? Part II: theory and intuition." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 384(2321):20240614.

Goyal, Tanushree. 2025. "Local political representation as a pathway to power: A natural experiment in India." *American Journal of Political Science* 69(2):516–530.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.

Harris, Adam, Kristen Kao and Ellen Lust. 2020. "Research Design for 'Levels Experiment'." Unpublished manuscript.

Henderson, John A., Logan Dancey, Stephen N. Goggin, Geoffrey Sheagley and Alexander G. Theodoridis. 2022. "Primary Divisions: How Voters Evaluate Policy and Non-policy Differences between Co-party Politicians." *Journal of Politics* 84(3):1760–1776.

Hobolt, Sara B. and Toni Rodon. 2020. "Cross-Cutting Issues and Electoral Choice: EU Issue Voting in the Aftermath of the Brexit Referendum." *Journal of European Public Policy* 27(2):227–245.

Holman, Mirya R., Jennifer L. Merolla and Elizabeth J. Zechmeister. 2016. "Terrorist Threat, Male Stereotypes, and Candidate Evaluations." *Political Research Quarterly* 69(1):134–147.

Hopkins, Daniel J. 2014. "Do Voting Criteria Differ across Levels of the U.S. Federal System? A Conjoint Experiment to Understand Voter Decision-Making.".
URL: *https://tessexperiments.org/study/hopkins365*

Horiuchi, Yusaku, Daniel M. Smith and Teppei Yamamoto. 2020. "Identifying voter preferences for politicians' personal attributes: a conjoint experiment in Japan." *Political Science Research and Methods* 8(1):75–91.

Horne, Will. 2020. "Do Voters Really Want Working Class Candidates?" Unpublished manuscript.

Kage, Rieko, Frances M. Rosenbluth and Seiki Tanaka. 2018. "What Explains Low Female Political Representation? Evidence from Survey Experiments in Japan." *Politics & Gender* pp. 1–25.

Kao, Kristen and Lindsay J. Benstead. 2021. "Female Electability in the Arab World: The Advantages of Intersectionality." *Comparative Politics* 53(3):427–464.

Kirkland, Patricia and Alexander Coppock. 2018. "Candidate Choice Without Party Labels:: New Insights from Conjoint Survey Experiments." *Political Behavior* 40:571–591.

Leeper, Thomas J. and Joshua Robison. 2020. "More Important, but for What Exactly? The Insignificant Role of Subjective Issue Importance in Vote Decisions." *Political Behavior* 42:239–259.

Lemi, Danielle Casarez. 2021. "Do Voters Prefer Just Any Descriptive Representative? The Case of Multiracial Candidates." *Perspectives on Politics* 19(4):1061–1081.

Magni, Gabriele and Andrew Reynolds. 2021. "Voter Preferences and the Political Underrepresentation of Minority Groups: Lesbian, Gay, and Transgender Candidates in Advanced Democracies." *The Journal of Politics* 83(4):1199–1215.

Martin, Nicole S. and Scott Blinder. 2021. "Biases at the Ballot Box: How Multiple Forms of Voter Discrimination Impede the Descriptive and Substantive Representation of Ethnic Minority Groups." *Political Behavior* 43:1487–1510.

Mo, Cecilia Hyunjung. 2015. "The Consequences of Explicit and Implicit Gender Attitudes and Candidate Quality in the Calculations of Voters." *Political Behavior* 37(2):357–395.

Ono, Yoshikuni and Barry C. Burden. 2019. "The Contingent Effects of Candidate Sex on Voter Choice"." *Political Behavior* 41(3):583–607.

Ono, Yoshikuni and Masahiro Yamada. 2020. "Do Voters Prefer Gender Stereotypic Candidates? Evidence from a Conjoint Survey Experiment in Japan." *Political Science Research and Methods* 8(3):477–492.

Peterson, Erik. 2017. "The Role of the Information Environment in Partisan Voting." *The Journal of Politics* 79(4):1191–1204.

Piliavin, Jane Allyn. 1987. "Age, Race, and Sex Similarity to Candidates and Voting Preference." *Journal of Applied Social Psychology* 17(4):351–368.

Saha, Sparsha and Ana Catalano Weeks. 2022. "Ambitious Women: Gender and Voter Perceptions of Candidate Ambition." *Political Behavior* 44:779–805.

Schaffner, Brian and Jon Green. 2019. "What Attributes Do Democratic Primary Voters Value?" *Data for Progress* .

URL: *https://www.dataforprogress.org/blog/2019/7/11/what-attributes-do-democratic-primary-voters-value*

Schuler, Paul. 2020. "Gender Bias and Position Salience: A Theory and Test of Variable Political Gender Bias in a Single-Party Regime." Unpublished manuscript.

Schwarz, Susanne and Alexander Coppock. 2022. "What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments." *The Journal of Politics* 84(2):655–668.

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." *Political Research Quarterly* 70:374–393.

Senninger, Roman and Daniel Bischof. 2021. "Do Voters Want Domestic Politicians to Scrutinize the European Union?" *Political Science Research and Methods* pp. 1–9.

Sigelman, Lee and Carol Sigelman. 1982. "Sexism, Racism, and Ageism in Voting Behavior: An Experimental Analysis." *Social Psychology Quarterly* 45(4):263–269.

Simas, Elizabeth N. and Doug Murdoch. 2019. "'I Didn't Lie, I Misspoke': Voters' Responses to Questionable Campaign Claims." *Journal of Experimental Political Science* 7(2):75–88.

Snagovsky, Feodor, Woo Chang Kang, Jillian Sheppard and Nicholas Biddle. 2020. "Does descriptive representation increase perceptions of legitimacy? Evidence from Australia." *Australian Journal of Political Science* 55(4):378–398.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112(3):525–541.

Tomz, Michael and Robert P. Van Houweling. 2016. "Political Repositioning: A Conjoint Analysis.".
URL: *https://tomz.people.stanford.edu/sites/g/files/sbiybj4711/f/tomz-vh-repositioning-conjoint-2016-04.pdf*

Visconti, Giancarlo. 2022. "Ideological Preferences after Natural Disasters: Evidence from Chile." *Political Behavior* 44(4):1985–2004.

Visconti, Giancarlo and Isabela Mares. 2020. "Voting for the Lesser Evil: Evidence from a Conjoint Experiment in Romania." *Political Science Research and Methods* 8(2):315–328.

Vivyan, Nick and Markus Wagner. 2015. "What Do Voters Want From Their Local MP?" *The Political Quarterly* 86(1):33–40.

Wuest, Reto and Jonas Pontusson. 2022. "Voter preferences as a source of descriptive (mis)representation by social class." *European Journal of Political Research* 61:398–419.