aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–11

DOI: 10.1177/0956797620972367
www.psychologicalscience.org/PS
SAGE

Asking People to Explain Complex Policies Does Not Increase Political Moderation: Three Preregistered Failures to Closely Replicate Fernbach, Rogers, Fox, and Sloman's (2013) Findings

Jarret T. Crawford and John Ruscio
Psychology Department, The College of New Jersey

Abstract

Fernbach et al. (2013) found that political extremism and partisan in-group favoritism can be reduced by asking people to provide mechanistic explanations for complex policies, thus making their lack of procedural-policy knowledge salient. Given the practical importance of these findings, we conducted two preregistered close replications of Fernbach et al.'s Experiment 2 (Replication 1a: $N = 306$; Replication 1b: $N = 405$) and preregistered close and conceptual replications of Fernbach et al.'s Experiment 3 (Replication 2: $N = 343$). None of the key effects were statistically significant, and only one survived a small-telescopes analysis. Although participants reported less policy understanding after providing mechanistic policy explanations, policy-position extremity and in-group favoritism were unaffected. That said, well-established findings that providing justifications for prior beliefs strengthens those beliefs, and well-established findings of in-group favoritism, were replicated. These findings suggest that providing mechanistic explanations increases people's recognition of their ignorance but is unlikely to increase their political moderation, at least under these conditions.

Keywords

replication, political psychology, political extremism, attitude change, open data, open materials, preregistered

Received 11/28/18; Revision accepted 9/19/20

Americans are increasingly politically polarized (Iyengar et al., 2012; Pew Research Center, 2017), are more motivated by negative views of the opposing party than positive views of their own (Abramowitz & Webster, 2016), and cannot agree on which issues facing the country are the most important (Pew Research Center, 2018). Along with increased polarization and partisanship, people's political identities have fused with other social identities (e.g., race and gender; Levendusky, 2009; Mason, 2018). Polarization and extremism can cripple democracies, which require engagement and compromise (Achen & Bartels, 2016; Fishkin, 2011). Unfortunately, some efforts to decrease polarization have even been counterproductive (e.g., exposure to counterattitudinal information on social media; Bail et al., 2018).

People generally possess low political knowledge (Carpini & Keeter, 1996) and often lack insight into their own ignorance (Kruger & Dunning, 1999). Recognizing this, Fernbach et al. (2013) reasoned that confronting people with their lack of procedural-policy knowledge would increase political moderation and decrease partisan out-group bias. They tested this prediction by asking participants to explain how a policy works—that is, to provide *mechanistic explanations* for policies. Consistent with this prediction, their findings showed that participants who provided mechanistic policy explanations

Corresponding Author:
Jarret T. Crawford, The College of New Jersey, Psychology Department
E-mail: crawford@tcnj.edu

reported decreased policy understanding and extremism on policies (Experiments 1 and 2). These effects were more pronounced relative to a condition in which participants were not confronted with their ignorance of policy procedure—that is, they simply provided reasons for their personal policy positions (Experiment 2). In Experiment 3, providing mechanistic explanations (compared with providing reasons) reduced intergroup bias; participants were less likely to donate to like-minded political organizations.

These findings have practical implications for reducing political extremism that have been recognized by the scientific community (Fernbach et al.'s article has been cited 43 times per year on average) and the public (e.g., see the headline, "Political extremism can be moderated by asking people a simple question"; Perry, 2013). It is therefore prudent to determine their replicability, especially in light of wider replicability concerns (Nelson et al., 2018) and low post hoc power estimates for the effect of mechanistic explanations on reducing extremity in the original article (.28 and .61 for Experiments 2 and 3, respectively). No systematic close replications of this influential work have been reported, although some conceptual replications suggest mixed support. In one sample (n = 296), Voelkel et al. (2018, supplemental materials, p. 8) replicated the effect of mechanistic explanations on understanding ($b = -0.38$, $p < .001$) but not on policy extremity ($b = 0.04$, $p = .675$); however, because there were intervening materials between the independent and dependent variables (e.g., a battery of intergroup attitudes), this study cannot be interpreted as a close replication. In another sample (n = 224), Johnson et al. (2016) reported effects of explanatory ability on understanding ($d = 0.52$) and extremity ($d = 0.33$), albeit with a very different manipulation than that used in the original experiments by Fernbach et al.

We therefore conducted two close replications of Fernbach et al.'s (2013) Experiment 2 (Replications 1a and 1b) and one close replication of Experiment 3 (Replication 2), which also included a conceptual replication following the original study materials. No additional samples beyond these three were collected. We did not attempt to replicate Experiment 1 because Experiment 2 itself was a replication and extension of Experiment 1. We used materials provided by the original authors (P. M. Fernbach, personal communication, February 18, 2018); when original materials were not available, we crafted them on the basis of their description in the original article. In March 2018, we randomly assigned each participant to Replication 1a or Replication 2 (preregistration: https://osf.io/fy4hz/); these participants were prevented from participating in Replication 1b (collected April 2018; preregistration: https://osf.io/8qz4f/). As in the original experiments, samples were drawn from Amazon's Mechanical Turk (MTurk). Materials and data for this project are available on the OSF (https://osf.io/zep2b/).

Replications 1a and 1b

In Experiment 2, Fernbach et al. found that policy understanding and political extremity decreased following mechanistic explanations. Critically, there was no decrease in extremity when participants provided reasons for their positions.

In Replication 1a, we failed to achieve the preregistered sample size of at least 2.5 times the original sample size (Simonsohn, 2015). Further, the original author could not provide the exact attention check used in the original experiment, so we developed our own. However, the failure rate (1%) in Replication 1a was sizably lower than in the original experiment (21%). We therefore conducted Replication 1b, which exceeded the preregistered sample size and included a more challenging attention check.

Method

Participants. In Experiment 2, Fernbach et al. recruited 141 participants, of whom 112 (79%) passed the attention filter and were included in their analysis. We recruited 306 participants for Replication 1a (2.17 times the original sample size) and 405 participants for Replication 1b (2.87 times the original sample size), of whom 302 (99%) and 377 (93%) passed the attention filter, respectively, and were included in the final analyses (Replication 1a: 51% female; 44% Democrat, 28% Republican, 27% independent, 1% other; mean age = 38 years; Replication 1b: 47% female; 42% Democrat, 25% Republican, 31% independent, 2% other; mean age = 38 years).

Materials and procedure. Replications 1a and 1b were nearly identical to each other; differences between the two are noted below.

Participants first provided demographic information (gender, age, education,[1] political party). They then provided their preexplanation position on the same six policies used in the original experiments: (a) imposing unilateral sanctions on Iran for its nuclear program, (b) raising the retirement age for Social Security, (c) transitioning to a single-payer health care system, (d) establishing a cap-and-trade system for carbon emissions, (e) instituting a national flat tax, and (f) implementing merit-based pay for teachers (1 = *strongly against*, 2 = *against*, 3 = *somewhat against*, 4 = *neither in favor nor against*, 5 = *somewhat in favor*, 6 = *in favor*, 7 = *strongly in favor*). Item order was randomized for each participant. Because Fernbach et al. did not specify a rationale for choosing these policies (e.g., relevance to national

politics at the time of data collection), we chose not to alter the set of policies.

Participants then read training instructions for how to quantify their level of policy understanding, using an unrelated policy issue (immigration reform) as an example. Afterward, they rated their preexplanation policy understanding of the six political issues (1 = *vague understanding*, 7 = *thorough understanding*), also presented in random order for each participant.

Each participant was then randomly assigned to either the reasons condition (Replication 1a: n = 171; Replication 1b: n = 205) or the mechanistic condition (Replication 1a: n = 130; Replication 1b: n = 172). In the reasons condition, participants were asked to list the reasons why they held their policy position. In the mechanistic condition, participants were asked to describe the chronological steps by which the policy is effected, making causal connections between each step. Regardless of condition, each participant was asked to consider two different issues (Iran sanctions followed by merit-based teacher pay, single-payer health care followed by Social Security retirement age, or cap-and-trade followed by flat tax). The Iran and merit-pay issues were not included in Fernbach et al.'s Experiment 2 because of a programming error. We included them in our analyses to follow the intent of the original protocol and because these issues were included in Fernbach et al.'s Experiment 1 analyses.

After providing their written response, participants rerated their level of understanding (i.e., postexplanation understanding rating) for that issue using the same 7-point scale as for the preexplanation understanding rating (e.g., "Please state your position on transitioning to a single-payer health care system"). They then rerated their position on that issue (i.e., postexplanation position rating) using the same 7-point scale as for the preexplanation position rating. Participants repeated these postexplanation understanding and position ratings after providing their written response to the second issue, as in the original experiment. Because the original authors were unable to provide exact rerating instructions, participants in the replications were simply asked to provide their issue positions without further elaborated instructions. Scatterplots of these before and after ratings for understanding and extremity are provided in Figures S2 and S3 in the Supplemental Material available online (these measures were highly positively correlated with one another, with rs ranging from .69 to .87).

In Replication 1a, after rating their postexplanation position on the second issue, participants completed the attention check, for which they were asked to select "agree" on a single 7-point agreement item to demonstrate that they were paying attention (1 = *strongly disagree*, 7 = *strongly agree*). We created a more challenging attention check in Replication 1b by embedding a similar attention-check item within a matrix of six other items at the end of the survey.[2]

Results

In Experiment 2, Fernbach et al. hypothesized that writing a mechanistic explanation of a particular policy would reduce participants' reported understanding of that policy and their position extremity, especially relative to a condition in which participants were asked to provide their reasons for holding their policy position. They first reported repeated measures analyses of variance on understanding and extremity (with extremity calculated as the absolute value of one's raw policy rating minus 4, the scale midpoint), with timing of judgment (before explanation vs. after explanation) and issue number (first issue vs. second issue) as within-subjects variables. They reported these analyses separately for the mechanistic and reasons conditions and then followed these analyses by reporting results from models in which condition was included as a between-subjects variable. We followed this analytic approach. In Figure 1, we report means for both understanding and extremity ratings, in both the mechanistic and reasons conditions, for the original Experiment 2, Replication 1a, and Replication 1b.

Mechanistic condition only.

Understanding. In Experiment 2, Fernbach et al. observed their critical main effect of timing: Reported understanding decreased following the mechanistic explanation. This replicated their own finding from their Experiment 1. They also observed an unexpected main effect of issue number: Participants reported understanding the first issue better than the second. There was no timing-by-issue-number interaction.

Table S1a in the Supplemental Material reports the replication findings. Replications 1a and 1b each replicated the main effect of timing: Participants' postexplanation understanding (Replication 1a: M = 3.61, SE = 0.14; Replication 1b: M = 3.62, SE = 0.12) was lower than their preexplanation understanding (Replication 1a: M = 3.91, SE = 0.14; Replication 1b: M = 3.88, SE = 0.12). Replications 1a and 1b each revealed a main effect of issue number but in the opposite direction from the original: Participants reported greater understanding of the second issue than the first. As in the original experiment, the interaction effects were not significant.

Extremity. In Experiment 2, Fernbach et al. observed their critical main effect of timing: Participants reported less extreme postexplanation positions than preexplanation

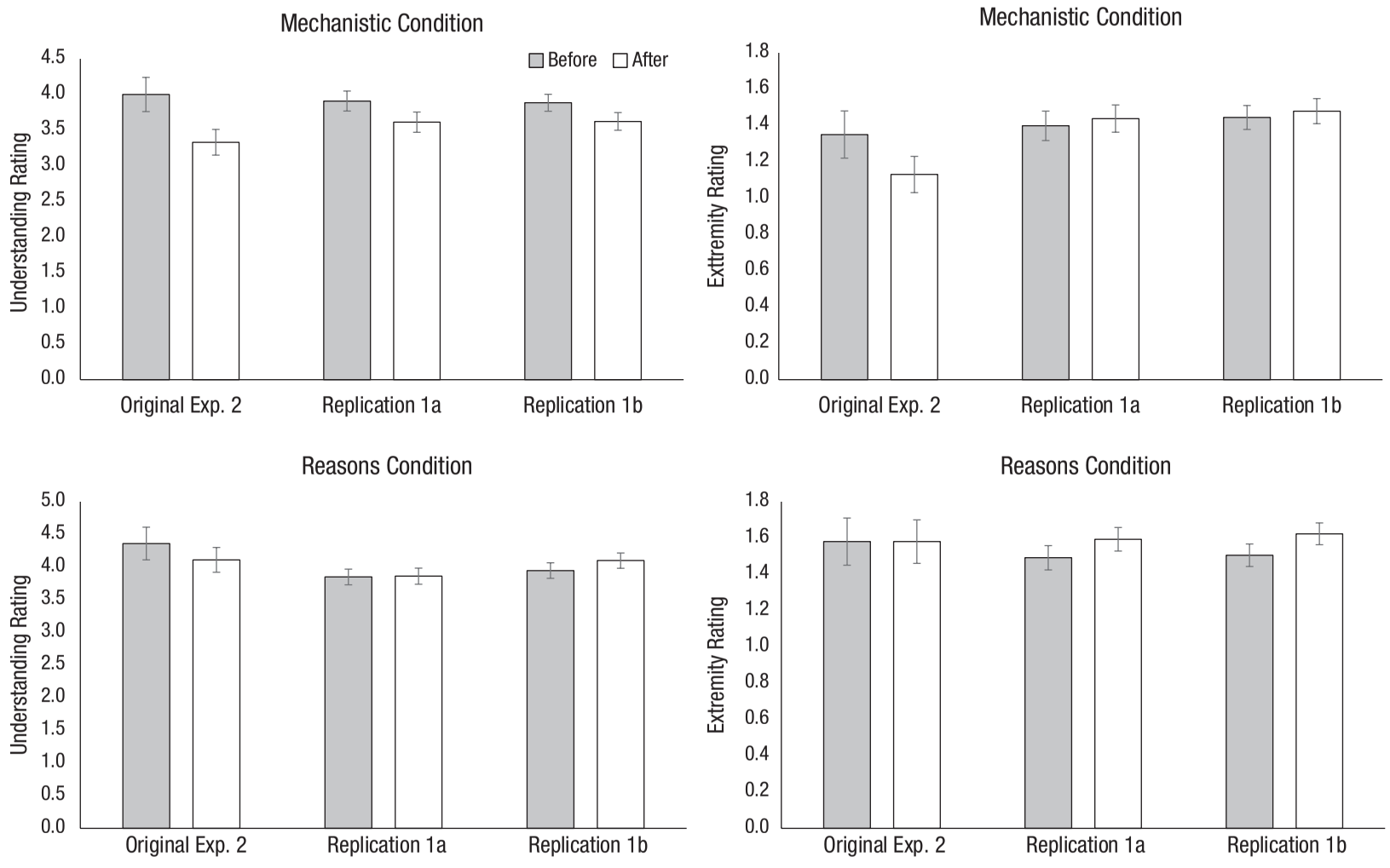

Fig. 1. Mean pre- and postexplanation ratings of understanding (left column) and extremity (right column) in the original Experiment 2 and Replications 1a and 1b, separately for the mechanistic and reasons conditions. Error bars indicate ±1 *SEM*.

positions. This replicated their own finding from their Experiment 1. They did not report the main effect of issue number or the timing-by-issue-number interaction.

Table S1b in the Supplemental Material shows that the critical main effect of timing was not replicated; there were no significant differences in preexplanation extremity ratings (Replication 1a: $M = 1.40$, $SE = 0.08$; Replication 1b: $M = 1.44$, $SE = 0.07$) relative to postexplanation extremity ratings (Replication 1a: $M = 1.44$, $SE = 0.08$; Replication 1b: $M = 1.48$, $SE = 0.07$). Again, there was the unexpected main effect of issue number, along with a small timing-by-issue-number interaction on extremity in both replications. Follow-up analyses revealed that there was a significant increase in extremity on the second issue (ps = .034 and .032 for Replications 1a and 1b, respectively). Although there was somewhat less extremity on the first issue, these differences were not significant (ps = .298 and .210 for Replications 1a and 1b, respectively).

Reasons condition only.

Understanding. Fernbach et al. reported a main effect of timing: Understanding decreased after participants provided reasons. Table S1c in the Supplemental Material shows that a main effect of timing was not observed in Replication 1a; there was no significant difference in preexplanation understanding ratings ($M = 3.85$, $SE = 0.12$) relative to postexplanation understanding ratings ($M = 3.86$, $SE = 0.12$). There was a significant main effect of timing in Replication 1b but in the opposite direction from the original: Postexplanation understanding ($M = 4.10$, $SE = 0.12$) was higher than preexplanation understanding ($M = 3.95$, $SE = 0.12$). The unexpected main effect of issue number was again observed. There was no timing-by-issue-number interaction.

Extremity. Fernbach et al. reported that the main effect of timing was not significant in the reasons condition. Inconsistent with the result of the original experiment, our findings showed main effects of timing, indicating that postexplanation extremity ratings were higher (Replication 1a: $M = 1.59$, $SE = 0.07$; Replication 1b: $M = 1.62$, $SE = 0.06$) than preexplanation extremity ratings (Replication 1a: $M = 1.49$, $SE = 0.07$; Replication 1b: $M = 1.51$, $SE = 0.06$; see Table S1d in the Supplemental Material). The effect of issue number was statistically significant in Replication 1a ($p < .001$) but not Replication 1b ($p = .094$). There were no significant timing-by-issue-number interactions in either replication.[3]

Mixed models. In the mixed models, Fernbach et al. reported significant timing-by-condition interactions on both understanding and extremity: Decreased postexplanation understanding and extremity ratings were more pronounced in the mechanistic condition than in the reasons condition. A timing-by-condition interaction on understanding emerged in both replications—Replication 1a: $F(1, 299) = 6.62$, $p = .011$, $\eta_p^2 = .02$; Replication 1b: $F(1, 374) = 15.31$, $p < .001$, $\eta_p^2 = .04$—suggesting that the manipulation did influence reported understanding as expected. However, the critical timing-by-condition interaction on extremity did not emerge in either replication—Replication 1a: $F(1, 293) = 0.66$, $p = .416$, $\eta_p^2 = .002$; Replication 1b: $F(1, 371) = 1.76$, $p = .185$, $\eta_p^2 = .005$. No other interactions (including the timing-by-issue-number-by-condition interactions) were significant in the replications.

Political extremity as an explanation for failure to replicate the original results? Given rising political polarization, one might wonder whether participants in the replications are simply more extreme than participants in the original experiment and whether this might account for the failure to replicate the original results. This does not appear to be the case, because the mean level of extremity in Fernbach et al.'s Experiment 2 ($M = 1.49$, on a scale from 0 to 4) and the mean levels of extremity in Replications 1a and 1b were nearly identical (Ms = 1.49 and 1.51, respectively, on the same scale).

Contrasting effects in the original and replication experiments. Asking participants to provide mechanistic policy explanations led them to report less policy understanding, especially relative to simply providing reasons for policy positions. However, this apparent recognition of policy ignorance did not translate into political moderation, because participants did not statistically significantly alter their issue positions after providing mechanistic explanations. We therefore failed to replicate the key finding from Fernbach et al.'s Experiments 1 and 2.

In considering the original and replication effects side by side (see Fig. 1), it does appear that the effects of mechanistic explanations on understanding were smaller in the replications than in the original Experiment 2. Although this might suggest that the effect on understanding was too weak to produce decreased extremity in our replications, there are several reasons to be skeptical of this interpretation. First, our samples were substantially larger than the original samples and produced highly consistent effect sizes across samples (η_p^2s = .07 and .05, respectively). Second, given that the original experiment was underpowered, the effect size ($\eta_p^2 = .31$) may have been overestimated. Indeed, the effect size (η_p^2) in Experiment 1 of the original article, which had a larger sample size than Experiment 2, was .15. Thus, we may have been more accurately estimating the effect size in these replications. Third, our findings do not suggest that a smaller effect of the independent variable on the manipulation check (i.e.,

understanding) necessarily led to a smaller effect on the dependent variable (i.e., extremity); if anything, extremity slightly increased in the mechanistic condition. Finally, Replications 1a and 1b each showed that providing reasons increased extremity, indicating that the extremity measure was sensitive to the writing task. This finding is inconsistent with Fernbach et al.'s findings but is consistent with other evidence that people become more extreme after being given the opportunity to justify their prior attitudes (e.g., Tesser et al., 1995).

Replication 2

With Replication 2, we attempted to closely reproduce Fernbach et al.'s Experiment 3 finding that providing mechanistic explanations reduces the likelihood of political donations to like-minded groups, and we looked to extend those findings to reducing intergroup biases (a conceptual replication).

Method

Participants. In Experiment 3, Fernbach et al. recruited 101 participants, of whom 92 (91%) passed the attention filter and were included in their final analysis. We recruited 343 participants (3.39 times the original sample size) for Replication 2, of whom 341 (99%) passed the attention filter and were included in the final analyses (55% female; 47% Democrat, 21% Republican, 29% independent, 3% other; mean age = 38 years).

Materials and procedures. Any discrepancies with the original experiment are identified below.

Participants first provided their policy positions using the identical issues from Replications 1a and 1b. Each participant was then randomly assigned to provide one of the following: a mechanistic explanation for cap-and-trade policy (n = 81), a mechanistic explanation for flat-tax policy (n = 78), reasons for their position on cap-and-trade policy (n = 82), or reasons for their position on flat-tax policy (n = 100).

Participants were then told that they would be given a $0.20 bonus payment that they could (a) donate to a group that advocated for the policy they were assigned, (b) donate to a group that advocated against the policy they were assigned, (c) keep for themselves, or (d) choose not to accept. The original authors could not provide the names of the original advocacy groups, so we used generic group names instead (e.g., "Donate the money to a group that advocates for establishing a cap and trade system for carbon emissions"). Decisions to donate to one's preferred advocacy group were coded as 1 (15% of participants), and all other choices were coded as 0 (81% kept the money for themselves, 3% turned the money down, 1% gave the money to an opposing group). Participants then encountered the attention check, which was identical to the one used in Replication 1a.

Following these original protocol materials, we presented materials for the conceptual replication. Specifically, participants completed feeling-thermometer ratings (0 = *very cold*, 50 = *neutral*, 100 = *very warm*) followed by social-distance ratings (1 = *very unwilling*, 6 = *very willing*) of people in favor of instituting a national flat tax, people against instituting a national flat tax, people in favor of a cap-and-trade system for carbon emissions, and people against a cap-and-trade system for carbon emissions.[4] Feeling-thermometer and social-distance ratings are the two most common measures of prejudice (Correll et al., 2010). The feeling-thermometer and social-distance ratings for each target were standardized on a scale ranging from 0 to 1 and were averaged together (all rs > .38, ps < .001). A difference score was created between the ratings of the two opposing groups; high scores on the measure of flat-tax bias indicated greater prejudice against flat-tax opponents than supporters, and high scores on the measure of cap-and-trade bias indicated greater prejudice against cap-and-trade opponents than supporters. Participants then provided demographic information.

Results

Close replication. As in the original experiment, we conducted a binary logistic regression, with donation decision as the outcome variable and condition (0 = reasons, 1 = mechanistic), extremism (mean centered), and their interaction as the independent variables. Fernbach et al. observed a significant interaction: People higher in extremism showed a reduced likelihood of donating to like-minded organizations in the mechanistic condition relative to the reasons condition, whereas condition did not influence donation likelihood among people lower in extremism. They did not report main effects. The interaction in the replication experiment was not significant, b = 0.30, SE = 0.32, Wald(1) = 0.87, p = .351. Thus, the original finding failed to closely replicate; people higher in extremism did not moderate their donation decisions toward like-minded groups more in the mechanistic than the reasons condition, which is inconsistent with the findings of the original experiment. There was a main effect of extremity, b = 0.63, SE = 0.16, Wald(1) = 16.49, p < .001: People higher in extremism donated to like-minded groups at a higher rate than people lower in extremism. There was no statistically significant main effect of condition, b = 0.23, SE = 0.31, Wald(1) = 0.51, p = .473.

Conceptual replication. The conceptual replication tested whether intergroup biases (in favor of like-minded over opposite-minded groups) would be weaker in the mechanistic condition relative to the reasons condition. To test this, we conducted hierarchical multiple regression analyses, regressing the measures of flat-tax bias and cap-and-trade bias (in two separate analyses) on condition, policy position (mean centered), and their interaction. Neither model revealed the anticipated interaction—flat tax: $b = -0.002$, $SE = 0.01$, $t(329) = -0.12$, $p = .900$; cap and trade: $b = -0.03$, $SE = 0.02$, $t(330) = -1.33$, $p = .185$; thus, our analyses failed to conceptually replicate the original finding. There were main effects of policy position: Higher support for flat-tax policy was associated with a greater bias in favor of flat-tax supporters over opponents, $b = 0.15$, $SE = 0.01$, $t(329) = 19.06$, $p < .001$, and higher support for cap-and-trade policy was associated with greater bias in favor of supporters of cap-and-trade policy over opponents of the policy, $b = 0.15$, $SE = 0.01$, $t(330) = 17.87$, $p < .001$. There were no statistically significant main effects of condition—flat tax: $b = 0.01$, $SE = 0.03$, $t(329) = 0.40$, $p = .687$; cap and trade: $b = 0.03$, $SE = 0.03$, $t(330) = 0.98$, $p = .327$.

Do Discrepancies in Attention-Check Failure Rates Explain the Replication Failures?

Fernbach et al. observed attention-check failure rates of 21% and 9% in Experiments 2 and 3, respectively, using an attention-check item embedded with other survey items (no attention check was included in Experiment 1). In Replications 1a and 2, failure rates were only about 1% using an attention check that was not well embedded with other survey items. Replication 1b used a well-embedded attention check, and we observed an attention-check failure rate of 7%, similar to rate in the original Experiment 3. To further identify and remove inattentive participants, we reviewed participants' open-ended responses to the reasons and mechanistic explanations (for details of the removal criteria, see the preregistration of this analysis: https://osf.io/a32cu). These exclusions, in addition to those based on the initial attention checks, yielded attention-check failure rates of 4.90%, 10.62%, and 4.40% in Replications 1a, 1b, and 2, respectively. The main findings remained unchanged under these more rigorous exclusion criteria as well as when no attention filters were used. It is unclear why our attention-check failure rates were generally lower than in the original experiments; one possibility is differences in participant quality: MTurk workers in the replication experiments needed at least a 95% human intelligence task (HIT) approval rating to participate, whereas the criterion for inclusion in the original experiments is unknown.

Would the Replication Effects Have Been Detectable in the Original Experiments?

We adapted Simonsohn's (2015) small-telescopes approach to evaluate the success of these replications. This approach examines whether the effect sizes observed in the replications would have been detectable in the original experiments. In other words, were the observed effects in the replications so small as to be undetectable using the original experiments' sample sizes? To evaluate this, we estimated the power of the within-subjects timing effect of mechanistic explanations on extremity in Replications 1a and 1b and the condition-by-extremism interaction effect on partisan in-group favoritism (i.e., political-donation decision) in Replication 2, given the original experiments' sample sizes. We calculated the 90% confidence intervals (CIs) around this estimate as well as the power associated with the upper bound of that CI. If the estimated power of the point estimate is lower than 33%, and especially if the estimated power at the upper bound of the 90% CI is less than 33%, then the original experiment would not have had adequate power to capture the effect observed in the replication. All of these analyses were performed using an empirical simulation method to construct CIs and estimate statistical power empirically (Ruscio, 2017).

Table 1 reports the original and replication sample sizes, effect-size estimates for each key effect and their corresponding power estimates, and effect-size estimates at the upper bound of the 90% CI and their corresponding power estimates. In no case did the power for the point estimate of the effect size exceed 33%, and the power estimate at the upper bound of the 90% CI exceeded 33% in only one case (62.1%; Replication 1b, on the understanding variable in the reasons condition). These findings suggest very weak statistical power to detect the effect sizes observed in the replications given the sample sizes in the original experiments. Figure 2 reports the effect sizes for the original and replication experiments.

General Discussion

Fernbach et al. reported an important and promising finding: In an era of heightened political polarization and animosity (e.g., Brandt & Crawford, 2019; Iyengar et al., 2012) and negative partisanship (e.g., Abramowitz & Webster, 2016), confronting people with their policy

Table 1. Results of Small-Telescopes Analyses

Original experiment	Replication experiment	Dependent variable	Original *N*	Replication *N*	Effect size	Statistical power	90% CI for effect size	Statistical power for upper bound of 90% CI
Exp. 2	Replication 1a: mechanistic condition	Extremity	47	130	0.05 (*d*)	5.60%	[–0.07, 0.17]	23.70%
Exp. 2	Replication 1b: mechanistic condition	Extremity	47	172	0.04 (*d*)	7.60%	[–0.04, 0.12]	24.70%
Exp. 2	Replication 1a: reasons condition	Understanding	65	171	0.01 (*d*)	5.00%	[–0.07, 0.08]	19.20%
Exp. 2	Replication 1b: reasons condition	Understanding	65	205	0.09 (*d*)	28.90%	[0.03, 0.15]	62.10%
Exp. 3	Replication 2	Donation	92	341	0.87 (χ^2)	1.90%	[0.01, 5.28]	8.70%

Note: In Replications 1a and 1b, the point estimate is Cohen's *d*. In Replication 2, the point estimate is Wald χ^2. In Replications 1a and 1b, dependent-samples *t* tests were used to compare pre- and postexplanation responses in the mechanistic condition only. In Replication 2, the condition-by-extremity interaction was tested across the entire sample. CI = confidence interval.

ignorance can not only lead to a recognition of that lack of understanding (Experiment 1 and 2) but also reduce political extremism (Experiment 1 and 2) and partisan in-group favoritism (Experiment 3).

Consistent with the original Experiment 2, Replications 1a and 1b showed that participants reported less policy understanding after providing mechanistic explanations. However, this recognition of their policy-procedure ignorance did not translate into more moderate issue positions, thus failing to replicate Experiment 2's critical finding. Participants' reported political extremism did, however, increase in the reasons condition in Replications 1a and 1b. Whereas this finding is inconsistent with Fernbach et al.'s data, it is consistent with other research demonstrating that extremism increases following opportunities to justify one's prior beliefs (Ross et al., 1977; Tesser et al., 1995). It also suggests that the political-extremity measure was sensitive to the writing task, casting doubt on explanations that suggest that the manipulations may not have been strong enough to affect the dependent variable or that participants were not attentive enough to the materials. In Replication 2, the mechanistic intervention had no effect on people's likelihood of donating to like-minded political groups (a close-replication failure), nor did it reduce partisan biases (a conceptual-replication failure). People donated to and expressed more positive attitudes toward like-minded over opposite-minded groups, consistent with evidence of partisan animosity (Brandt & Crawford, 2019).

Contrasting demographic characteristics in the original and replication samples does not help explain the replication failures. Figure S1 in the Supplemental Material shows that participants in these five samples were fairly similar in terms of gender and age. There were somewhat more political independents in Experiment 2 than in the replications. That said, the political composition of Experiment 3 is unknown, and the estimates provided in the original and replication experiments do not deviate much from estimates from MTurk and nationally representative samples (Levay et al., 2016).

These replication failures cannot be easily attributed to differences in political extremity between data-collection periods, as participants' average extremism ratings in Replications 1a and 1b were equivalent to those in Fernbach et al.'s Experiment 2. That said, given that contextual sensitivity is associated with lower replicability rates (Van Bavel et al., 2016), we cannot rule out the possibility that changing social contexts may partly explain these replication failures. It is possible that Fernbach et al.'s hypothesized effects were observable under the contextual conditions at the time but not at the time of our replication. Such moderating factors should be explored in future work.

There was some variation in participants' initial polarization on the policy issues in our samples (e.g., roughly 35% of participants across experiments gave extreme positions on single-payer health insurance, whereas only about 16% did so for merit-based teacher pay), but issue polarization was neither intentionally selected nor systematically varied in the original or the replication experiments. This moderator might be considered in future work.

The original and replication experiments were conducted in the United States, but political polarization is not strictly an American phenomenon (Pennycook et al., 2021). These processes should be explored in other national contexts in future work. Neither the

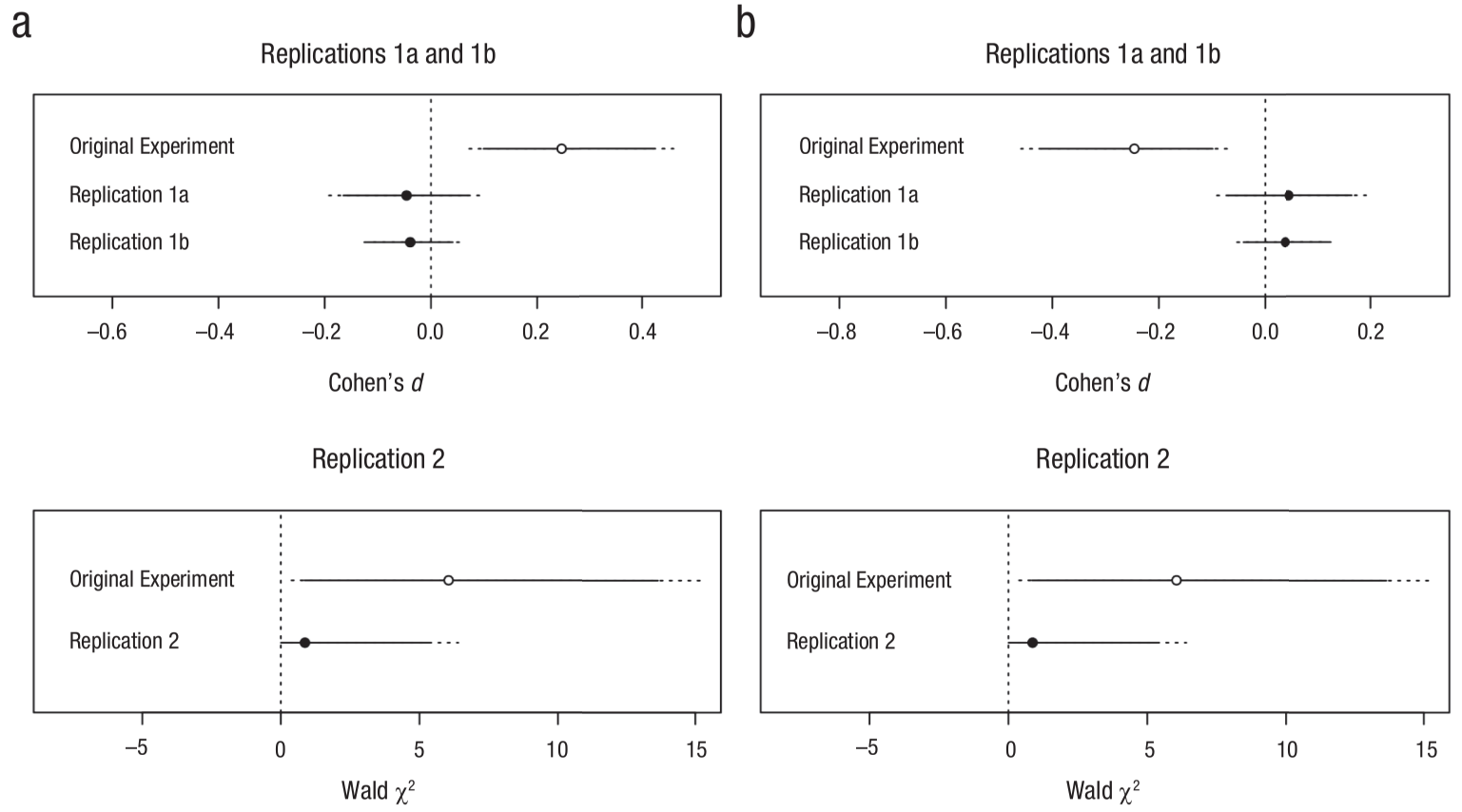

Fig. 2. Effects of mechanistic explanations on political extremism in the original and replication experiments. For Replications 1a and 1b, this figure reflects the within-subjects effect of timing on extremity in the mechanistic condition. For Replication 2, it reflects the condition-by-extremity interaction on partisan in-group favoritism. Solid horizontal lines represent 90% confidence intervals (CIs); dashed extensions of those horizontal lines represent 95% CIs. Dashed vertical lines represent the zero point for the effect-size test statistics; CIs that cross this line for Cohen's *d* (Replications 1a and 1b) indicate failure to reject the null hypothesis. For Wald χ^2 (Replication 2), CIs cannot include negative values because χ^2 itself cannot be a negative value.

original nor the replication experiments specifically targeted extremist participants; targeting such populations to enhance not only statistical power but also the relevance of the findings to typical political discourse could be a consideration for future work.

Given the importance of reducing political extremism, the failure of these replications to produce evidence that mechanistic explanations for policy positions reduce political extremism or partisan in-group favoritism is disappointing. These replication failures do not necessarily demonstrate that the effect does not exist, but they do suggest that it is not as robust as previously reported and that it requires cautious and precise further exploration.

Transparency

Action Editor: D. Stephen Lindsay
Editor: D. Stephen Lindsay
Author Contributions

J. T. Crawford designed the experiments, conducted most of the analyses, and drafted the manuscript. J. Ruscio conducted some analyses and helped edit the manuscript. Both authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

All data have been made publicly available via OSF and can be accessed at https://osf.io/zep2b. The design and analysis plans for the replications of Fernbach et al.'s (2013) Experiments 2 and 3 were preregistered at https://osf.io/fy4hz/. The preregistration for an additional replication of Fernbach et al.'s Experiment 2 is available at https://osf.io/8qz4f/. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jarret T. Crawford https://orcid.org/0000-0001-7885-0759

Acknowledgments

We thank Stephanie Mallinas for her comments on a draft of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797620972367

Notes

1. Education was assessed in the replication but not in the original experiments.
2. These items asked participants to rate how liberal or conservative each policy position was (identical to Replication 2).
3. In their reasons-condition-only analyses, Fernbach et al. did not report main effects of issue number or timing-by-issue-number interactions. We report them for these replications to provide the full set of analyses.
4. Participants completed similar ratings of liberals and conservatives and were asked whether they saw flat tax and cap-and-trade policies as liberal or conservative. These variables were not used in the analyses.

References

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of U.S. elections in the 21st century. *Electoral Studies*, *41*, 12–22.

Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, USA*, *115*(37), 9216–9221.

Brandt, M. J., & Crawford, J. T. (2019). Studying a heterogeneous array of target groups can help us understand prejudice. *Current Directions in Psychological Science*, *28*(3), 292–298.

Carpini, M. X. D., & Keeter, S. (1996). *What Americans know about politics and why it matters*. Yale University Press.

Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Measuring prejudice, stereotypes and discrimination. In J. F. Dovidio, M. Hewstone, P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping and discrimination* (pp. 45–62). SAGE.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946.

Fishkin, J. S. (2011). *When the people speak: Deliberative democracy and public consultation*. Oxford University Press.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431.

Johnson, D. R., Murphy, M. P., & Messer, R. M. (2016). Reflecting on explanatory ability: A mechanism for detecting gaps in causal knowledge. *Journal of Experimental Psychology: General*, *145*, 573–588.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *SAGE Open*, *6*(1). https://doi.org/10.1177/2158244016636433

Levendusky, M. (2009). *The partisan sort: How liberals became Democrats and conservatives became Republicans*. University of Chicago Press.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534.

Pennycook, G., McPhetres, J., Bago, B., & Rand, D. G. (2021). *Beliefs about COVID-19 in Canada, the U.K., and the U.S.A.: A novel test of political polarization and motivated reasoning*. PsyArXiv. https://doi.org/10.31234/osf.io/zhjkp

Perry, S. (2013, June 7). Political extremism can be moderated by asking people a simple question. *MinnPost*. https://www.minnpost.com/second-opinion/2013/06/political-extremism-can-be-moderated-asking-people-simple-question/

Pew Research Center. (2017). *The shift in the American public's political values: Political polarization, 1994-2017*. http://www.people-press.org/interactives/political-polarization-1994-2017/

Pew Research Center. (2018). *Little partisan agreement on the pressing problems facing the U.S*. http://www.people-press.org/2018/10/15/little-partisan-agreement-on-the-pressing-problems-facing-the-u-s/

Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology*, *35*, 817–829.

Ruscio, J. (2017). Performing "small telescopes" analysis by simulation: Empirically estimating statistical power and constructing confidence intervals [Manuscript in preparation]. Psychology, The College of New Jersey.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Tesser, A., Martin, L., & Mendolia, M. (1995). The impact of thought on attitude extremity and attitude-behavior consistency. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 73–92). Erlbaum.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, USA*, *113*(23), 6454–6459.

Voelkel, J. G., Brandt, M. J., & Colombo, M. (2018). I know that I know nothing: Can puncturing the illusion of explanatory depth overcome the relationship between attitudinal dissimilarity and prejudice? *Comprehensive Results in Social Psychology*, *3*, 56–78.