Effects of inoculation interventions on impressions of mock political candidates

Michael S. Cohen, Jean Decety
University of Chicago

Author note: We thank Aseem Gidwani for assisting with this work. Preregistrations, raw data, and analysis code from these studies will be available via OSF at https://osf.io/5vykp.

*Address correspondence to:
Michael S. Cohen or Jean Decety
Department of Psychology
University of Chicago
5848 S. University Ave.
Chicago, IL 60637
E-mail: mscohen@uchicago.edu (M.C.), decety@uchicago.edu (J.D.)

Abstract
Inoculation interventions have been a major focus of recent research attempting to combat the influence of disinformation in our society. However, no prior work has examined whether these interventions specifically reduce the impact of unproven or false accusations against politicians, which is a major vector by which disinformation can influence voting behavior in democracies. Here we demonstrate, through 11 studies, that interventions designed to increase awareness of manipulative messaging, such as the use of emotionally-charged language, can reduce the impact of derogatory accusations on impressions of novel mock political candidates. Individuals who viewed video or gamified inoculations showed a reduced impact of accusations on subjective ratings for all candidates about whom accusations of misconduct were made, relative to those who did not receive the inoculation. These effects were strongest when the truth status was ambiguous, but there is also some beneficial impact even when an accusation is clearly refuted, i.e., a reduction in Continued Influence Effects. Overall, these findings indicate that inoculation interventions that train citizens to be aware of manipulative techniques can effectively reduce the impact of unproven derogatory accusations against politicians.

Keywords: misinformation, inoculation, attitudes, political behavior, impression formation

Introduction

Misinformation and disinformation are not new phenomena. However, the rise of social media has dramatically amplified their spread, reach, and influence, producing deleterious effects that can potentially undermine the trust in common facts on which democratic societies depend. One significant social harm of disinformation to society, particularly prevalent in the contemporary environment, is the use of false accusations of misconduct to create negative impressions of political candidates. Empirical evidence shows that candidates targeted by such accusations are evaluated more negatively than comparable candidates who are not, even after the accusation has been fully corrected with factual information [1-2]. In zero-sum electoral competition, this dynamic confers an unfair advantage on candidates who are willing to disseminate falsehoods and lies about their opponents, or who tacitly allow their allies to do so. Although correcting false information can mitigate some effects of misinformation, false claims often continue to exert residual influence after correction. Moreover, fact-checking is a limited and increasingly contested resource, particularly as fact-checkers themselves become targets of coordinated attacks by disinformation actors. These constraints underscore the urgent need to develop interventions that extend beyond traditional fact-checking to more effectively reduce the impact of disinformation.

An influential innovation in recent years, which shows promise towards decreasing the impact of misinformation on society, is inoculation interventions [3-4]. These interventions are designed to make people preemptively aware of techniques that may be used to mislead them online or in other media, with the goal of stimulating critical thinking and skepticism against false information when it is encountered in the real world. While the underlying theory was developed decades ago [5], the methods have advanced substantially in recent years. Some key developments have included the use of "broad spectrum" interventions [e.g., 6], and the adoption of scalable gamified and video interventions with high production values [7-8]. This approach has proved promising as a way to counter misinformation in a variety of content domains.

Inoculation interventions are typically deemed successful when they increase an individual's ability to recognize and refrain from sharing false information without producing a reduction in belief for true information. For instance, validation for the first gamified inoculation approach was obtained by demonstrating that the intervention reduced reliability ratings of mock social media posts employing techniques referenced in the inoculation game, while having little effect on reliability ratings of true content [7]. Similarly, inoculation videos were validated on a range of measures, comparing mock social media posts designed to be either manipulative or neutral on reliability ratings, confidence in judgments of reliability, trustworthiness ratings, and willingness to share [8]. Again, the inoculation videos were largely effective across these measures. Similar measures were also used to evaluate a gamified intervention targeting COVID misinformation [9]. Some have argued, based on signal detection theory analyses, that inoculation primarily shifts the response criterion rather than enhancing the ability to

discriminate true from false information, reflecting a general increase in skepticism rather than improved discernment [10]. A subsequent meta-analysis using more data and more sophisticated statistical methods found that inoculation improves discernment without consistently changing response criterion [11]. Still, the focus remains on explicit accuracy judgments, and in some cases on willingness to share content, rather than on attitudes.

It remains unclear whether reducing explicit belief in disinformation also attenuates its influence on attitudes and behaviors toward its targets. Work on continued influence effects (CIEs) demonstrates that even when individuals are aware that presented information is false, the refuted information can continue to affect their judgments. CIEs were originally demonstrated in the context of causal reasoning about events; for instance, participants initially informed that flammable chemicals were present at a site of a fire continue to cite these chemicals as a cause of the fire even after being told that the initial report was erroneous and that no such chemicals were not present [12]. This work has largely focused on cognitive mechanisms, such as the persistence of retracted information within causal models, or better retrieval of the initial misinformation relative to its correction [13-14]. Still, CIEs can also occur in evaluation of other people, especially in political contexts [1-2], and effects persist even when individuals later remembered the retraction [2]. Thus, it is possible that following an inoculation intervention, individuals become aware that they are being exposed to false information, yet this awareness alone may be insufficient to eliminate the downstream effects of misinformation on subsequent judgments.

A few studies have tested outcome measures centered on behavioral intentions rather than beliefs, though only in a limited range of content domains. For instance, inoculation seems to partially counteract effects of false claims of disagreement among scientists about whether humans cause climate change [6, 15]. Specifically, when inoculation preceded false information, estimates of scientific consensus were higher and thus more accurate. There was also evidence for increased accuracy in people's understanding of the causes of climate change, as well as increased support for policies that fight climate change. Similarly, a gamified intervention intended to counter disinformation contributing to vaccine hesitancy in Africa led to increases in intention to get vaccinated, and to generally more positive attitudes towards vaccines [16]. At the same time, one study tested a range of inoculation approaches and found no substantial effects of these interventions in countering the impact of climate-related disinformation on the (in)accuracy of general beliefs about climate change [17]. Thus, more data is needed to clarify conditions under which inoculation affects attitudes.

A recently-proposed framework [18] is useful for considering the relationship between beliefs, attitudes, and behavioral intentions. This proposal emphasizes the role of belief-to-behavior inferences, which can be made at the time of initial exposure and stored in memory for later use. It is also noted that attitudes potentially mediate the relationship between beliefs and behavioral intentions [19-20]. Thus, a person who reads a narrative about misconduct by a politician may form a negative attitude and subsequently conclude

that they would not want to vote for that politician. When additional information is made available indicating that the misconduct accusation was false, not only does the false belief need to be revised, but so too should the attitudes and planned behaviors that followed from the false belief. Failures to do so could account for the observed CIEs in attitudes and voting intentions about hypothetical political candidates [1-2].

Another key question in the present study is whether inoculation interventions make corrections more effective. If participants are skeptical of accusations using persuasive techniques that are often misleading (e.g., emotional language), they may form a weaker belief-to-behavior inference at the outset. Thus, they will be more open to changing these inferences after a correction, enhancing the correction's effectiveness in returning attitudes and behavioral choices to what they would have been without the accusation.

Here, we focus on subjective attitudes toward hypothetical political candidates using feeling thermometer ratings. Following [2], *corrected accusation* candidates were defined as hypothetical candidates about whom a social media post alleging misconduct was presented, immediately followed by a second post, ostensibly from a journalistic source, refuting that accusation. CIEs were operationalized as the extent to which ratings were reliably lower for candidates with corrected accusations than for candidates about whom topically matched neutral stimuli were presented (*no accusation* candidates). Some candidates were presented with *uncorrected accusations*, i.e., an accusation followed by a control stimulus instead of a refutation. These candidates were generally given lower ratings than either the corrected accusation candidates or the no accusation candidates, which is normatively appropriate given that such accusations are at best unconfirmed and at worst plausibly true.

The extent to which an inoculation intervention reduces the rating difference between candidates facing *uncorrected accusations* and those facing no accusation indexes the degree to which inoculation reduces the impact of accusations on attitudes. If inoculation interventions also reduce the rating difference between *corrected accusation* candidates and those with no accusation, this would suggest that the preemptive skepticism induced by inoculation also allows factual corrections to more effectively modify attitudes. The two outcome measures of primary interest were ratings collected at an *immediate* time point, i.e., right after the social media posts were presented, as well as at a *delayed* time point, typically 15-20 minutes later. The latter measure tested the persistence of effects from inoculation interventions. In a subset of experiments, ratings were also obtained in a follow-up session two days after the original protocol to determine the longer-term persistence of effects of inoculation. In one large experiment, we additionally measured recognition memory at a two-day delay to evaluate mechanistic hypotheses motivated by prior work using this paradigm [2].

Data were collected across three large pre-registered online experiments ($n \approx 1500$-1600 each), 7 smaller online pilot studies (typical $n \approx 200$-300 each), and one pilot study

conducted with an undergraduate participant pool (n = 70), all employing similar protocols. Preregistered analyses were designed to test the comparisons described above. The final analyses deviated from the preregistration due to substantial unanticipated and unexplained heterogeneity between studies for some key measures, which limited the interpretability of any single experiment. It was therefore decided post hoc to emphasize hypothesis tests based on data aggregated across all 11 studies using internal meta-analyses. While this analytical strategy was not pre-registered, all experiments were included in the meta-analyses, leading to unbiased estimates of true effect sizes.

Method

Participants

Participants were recruited online from the CloudResearch Connect platform or other platforms as noted. Sample sizes for each study, compensation, and recruitment platform are noted in Supplemental Table 1. All studies that tested inoculation manipulations are included here. For the first 6 pilot experiments, all participants were included unless they failed a basic attention check. For the final two pilot studies, and for all three studies with larger samples, participants were also screened based on providing a correct answer to a multiple-choice question about the control video (which all participants viewed). The total aggregated sample included in the reported analyses were 6255 participants for measures based on immediate ratings, and 5745 participants for measures based on delayed ratings. The study design was reviewed and declared exempt from further review by the University of Chicago IRB.

Procedure

After providing informed consent, participants responded to basic attention checks: choosing which number is odd (11, 24, or 1120) and selecting a picture of a cat vs. a dog and a mouse. Then, following Cohen et al. (2024), each participant saw one-page neutral bios for each of the 27 hypothetical political candidates, with AI-generated faces and "bumper stickers" with candidate names. Immediately after each bio, participants provided an *initial rating* on a feeling thermometer scale ranging from 0 to 100, where 0 is very negative and 100 is very positive. This initial rating constituted a baseline rating for each candidate, with the same rating procedure repeated again later.

Inoculation intervention

Inoculation interventions were shown next (see Supplemental Table 2 for additional details). For the three large experiments and pilot studies 4, 5, 6, 7, and 8, the inoculation intervention was the emotional language video validated in [8], available at https://inoculation.science/inoculation-videos/emotional-language/ and at https://www.youtube.com/watch?v=ER64qa_qnWg . The control video matched to this inoculation video was a YouTube video about freezer burn (https://www.youtube.com/watch?v=fPEtOaGTZ0s). For Pilots 1, 2, and 3, the inoculation intervention was the "Bad News Game" validated in [7] (https://www.getbadnews.com/en), while a Tetris game (https://www.lumpty.com/amusements/Games/Tetris/tetris.html) was the control.

In the three large experiments, and in pilot studies 7 and 8, all participants watched the control video first and answered two questions about it. One question was easy (multiple-choice with a single answer) and one was hard (requiring the correct two out of five options to be selected). The easy screening question about the control video was used as a screening measure to exclude inattentive participants in these studies. Participants in the inoculation condition then viewed the inoculation video and answered one easy and one hard question about that video.

In Pilots 1, 2, 3, 4, 5, and 6, participants saw only one video/game (either inoculation or control). In Pilots 4 and 5, participants answered one hard question about the video they watched, which required selecting multiple correct answers. In Pilot 6, participants answered two multiple choice questions (each with a single correct answer) about the video they watched.

Primary Stimuli

After the inoculation intervention, the primary stimuli were presented. Following [2], the 27 candidates were evenly divided between *corrected accusation*, *uncorrected accusation,* and *no accusation* conditions, with assignment of candidate to condition counterbalanced. For each stimulus in the *corrected accusation* or *uncorrected accusation* conditions, Post 1 was a Facebook post accusing the politician of misconduct, with a linked article below the text (which was only a preview and could not actually be clicked). In the *corrected accusation* condition, Post 2 was a refutation framed as an investigation putatively from a journalistic source clearly indicating that the rumored accusation was not correct, and briefly explaining why, as well as showing a preview to a hypothetical article providing more detail. In the *uncorrected accusation* condition, Post 2 was a matched control stimulus that was similar to the refutation in terms of topic and structure but did not address the accusation. In the *no accusation* condition, Post 1 was a control post matched to the accusation in terms of topic but not accusing the politician of any sort of misconduct, while Post 2 was the same neutral post as used for the uncorrected accusation condition.

Participants provided immediate feeling thermometer ratings about each candidate just after reading Post 1 and Post 2 for each candidate. After all posts had been presented and immediate ratings were provided, in all three full experiments and in pilot experiments 4, 5, 7, and 8, participants in the inoculation condition viewed a 30 second “booster inoculation” video. This was an abbreviated version of the original inoculation message, available online at: https://www.youtube.com/watch?v=qE8Q6Fj_2Rg . Control participants did not see a video at this time. In full experiments 1 and 2, the MIST-20 [21], a test of one’s ability to explicitly detect headlines that are likely to be false, was administered next. Finally, in all three full experiments, participants provided another round of ratings after a delay. Here, only the “bumper sticker” with the candidate’s name and face was presented, and participants needed to again rate the candidate on a 0-100 feeling thermometer scale.

In full experiment 3, and in pilot studies 2 and 5, participants were asked to return for a second session two days after the initial session. Here, participants again provided feeling thermometer judgments in response to the "bumper sticker" with each candidate's name and face. The MIST-20 was also administered at this follow-up session (and not in the first session) in full experiment 3, as well as in pilot study 2. Finally, in full experiment 3, participants completed a memory test in this follow-up session. The procedure was similar to that used in prior work [2]. Specifically, all four stimuli for each candidate were shown, two of which were presented earlier in the main part of the study and two of which were not. The portion of the memory test that was used for the analyses reported here was the first half of the test; here, stimuli were configured so that only one of the accusation or the matched Post 1 control stimulus were presented for a given candidate, and only one of the refutation or the matched Post 2 control stimulus were presented for a given candidate, with assignment to half counterbalanced across participants. Hit rates and false alarm rates were computed for each of the four types of stimuli, and a d' score was then calculated to represent overall sensitivity for each item type.

Additional Measures

The following additional measures were included in some pilot studies but are not discussed here (see Supplemental Table 2): epistemic beliefs questionnaire [22], digital literacy measure [2], measures of affective polarization (a partisan feeling thermometer rating and a 3-item dictator game measure), the CTSQ, a self-report measure of intuitive vs. analytic thinking [23], 24 news headlines (50% true, 50% false; 1/3 Democrat-favoring, 1/3 Republican-favoring, 1/3 politically neutral), which people were asked if they would share this on their personal social media feed, a Javascript version of the perceptual metacognition task [24], dangerous world beliefs [25] and moral absolutism (an expanded version of the scale introduced in [26]).

In pilot studies 3-8, mock vote choices (following [2]) were collected just prior to the delayed ratings, at the end of the primary experimental session. This task was removed in the three full-sample studies to reduce the study duration, and because it was not as amenable to the linear mixed-model analysis approach that will be used as the primary analyses. In pilot study 2 and pilot study 5, long-delay choices were collected alongside long-delay ratings in the second session.

Analysis

Linear mixed effects models were used as primary analyses of feeling thermometer ratings for all 11 experiments. The first predictor variable (varied within-participants) is item type, modeled using dummy codes for *corrected accusation* and *uncorrected accusation* stimuli, relative to the baseline *no accusation* condition. The second predictor variable (varied between-participants) is inoculation condition, with the inoculation treatment group modeled relative to the control condition as baseline. Interactions between inoculation condition and both corrected accusation and uncorrected accusation dummy variables were also modeled. Initial ratings, made prior to the presentation of accusation or control stimuli, were modeled as a control variable. Finally, random

intercepts were modeled for each participant. When combining data across studies, the coefficient estimates and standard errors were entered into a random effects meta-analysis computed using REML, via the *rma* function in the *metafor* package in R with inverse-variance weighting applied (based on 1 / squared standard error).

Results

Immediate Ratings

First, candidate ratings provided immediately after people read the accusation for a given candidate are examined. Main effects indicating a reduction in ratings for candidates targeted by *uncorrected accusations* (see Figure 1) were highly significant in full experiment 1 (b = -22.65, t = -73.37, p < .0001), full experiment 2 (b = -24.05, t = -83.43, p < .0001), and full experiment 3 (b = -22.97, t = -81.68, p < .0001). A meta-analysis combining across all 11 experiments similarly showed a significant effect, as the weighted average unstandardized magnitude of the change in ratings for candidates with uncorrected accusations was -23.03 points (95% CI bounds = -23.97, -22.09), z = -43.11, p < .0001, on the 0-100 rating scale (Supplemental Figure 1A), with significant heterogeneity between studies, I^2 = 81.76%, Q = 42.44, p < .0001. Similarly, a significant Continued Influence Effect (i.e., reduced ratings for *corrected accusation* candidates vs. baseline) was apparent (see Figure 1) in full experiment 1 (b = -5.25, t = -17.01, p < .0001), full experiment 2 (b = -6.23, t = -21.61, p < .0001), and full experiment 3 (b = -6.27, t = -22.30, p < .0001). A meta-analysis combining across all 11 studies showed this effect with a weighted average unstandardized magnitude of -6.07 points (95% CI bounds = -6.98, -5.16), z = -13.04, p < .0001 (Supplemental Figure 1B). There was significant heterogeneity between studies here as well, I^2 = 80.82%, Q = 35.78, p < .0001. Ultimately, consistent with prior studies using these stimuli, both uncorrected and corrected accusations had a substantial negative effect on candidate ratings before considering effects of the inoculation intervention.

Next, we examine interaction terms between stimulus condition and inoculation condition to determine whether inoculation lessened the impact of accusation stimuli. Indeed, for uncorrected accusations, the positive effect of inoculation is significant in full experiment 1 (b = 1.39, t = 3.21, p = .001), full experiment 2 (b = 3.52, t = 8.53, p < .0001), and full experiment 3 (b = 2.69, t = 6.48, p < .001). A meta-analysis showed a significant weighted average unstandardized effect of 3.10 points (95% CI bounds = 1.42, 4.78), z = 3.61, p = .0003, and significant heterogeneity between studies, I^2 = 87.76%, Q = 59.88, p < .0001 (Figure 2). For corrected accusation stimuli, the effect of inoculation on corrected accusation stimuli was significant in full study 2 (b = 1.52, t = 3.91, p < .001) but not in full study 1 (b = 0.44, t = 1.02, p = .31) or full study 3 (b = 0.60, t = 1.43, p = .15). Across all experiments, however, a meta-analysis showed a small but statistically significant aggregate effect of inoculation on ratings, with a weighted average unstandardized effect of 1.03 points (95% CI bounds = 0.10, 1.96), z = 2.16, p = .031 (see Figure 2B), and significant heterogeneity, I^2 = 65.42%, Q = 24.47, p = .0065. Because of the high level of design similarity between experiments, the aggregate effect should be considered the best estimate of the true effect size, rather than the effect in any single experiment.

Beyond the reduced impact of accusations on immediate ratings after inoculation, it is meaningful to consider whether inoculation can entirely eliminate impacts of either uncorrected or corrected accusations on candidate ratings. This was not the case. Specifically, for participants in the inoculation condition, the weighted average unstandardized effect of uncorrected accusations, relative to no accusation candidates, was -19.92 points (95% CI bounds = -21.54, -18.30), z = -24.09, p < .0001, with significant heterogeneity, I^2 = 93.20%, Q = 117.60, p < .0001. Similarly, the weighted average unstandardized effect of corrected accusations was -5.01 points (95% CI bounds = -5.40, -4.62), z = -25.17, p < .0001, with no significant heterogeneity, I^2 = 27.62%, Q = 11.37, p = .33. The latter result indicates that Continued Influence Effects persist, despite being weakened somewhat, even when participants have been trained to be skeptical of misleading persuasive techniques.

Delayed Ratings

Candidate feeling thermometer ratings were also measured after a short delay but within the same experimental session. Here, participants were shown only the candidate's name, face, and banner, and had to again provide a rating on a 0-100 scale. Main effects (in individuals who did not receive the inoculation) show that uncorrected accusation stimuli were given lower ratings, relative to no accusation stimuli, in full experiment 1 (b = -4.90, t = -18.37, p < .0001), full experiment 2 (b = -5.12, t = -20.60, p < .0001), and full experiment 3 (b = -5.36, t = -22.72, p < .0001). A meta-analysis showed a weighted average unstandardized effect of -5.09 points (95% CI bounds = -5.34, -4.83), z = -39.28, p < .0001 (see Figure 3A), and no heterogeneity between studies, I^2 = 0.00%, Q = 4.85, p = .77. Corrected accusation stimuli were also given lower ratings relative to no accusation stimuli, and these effects were significant in full experiment 1 (b = -1.74, t = -6.54, p < .0001), full experiment 2 (b = -1.97, t = -7.93, p < .0001), and full experiment 3 (b = -2.10, t = -8.91, p < .0001). A meta-analysis found a significant effect here as well, with a weighted average unstandardized effect of -1.82 points (95% CI bounds = -2.08, -1.57), z = -14.05, p < .0001 (see Figure 3B), and no heterogeneity between studies, I^2 = 0.00%, Q = 6.04, p = .64. Thus, it is clear that negative effects of both uncorrected and corrected accusations on ratings of candidates are maintained after a short delay.

Finally, effects of inoculation on delayed ratings were examined. Within individual large-sample studies, inoculation significantly reduced the impact of *uncorrected accusations* in full study 2 (b = 1.14, t = 3.21, p = .001), with a marginal effect in full study 3 (b = 0.66, t = 1.90, p = .058) and no significant effect in full study 1 (b = 0.27, t = 0.72, p = 0.47). Still, the meta-analysis revealed that across all experiments, the inoculation intervention did reduce the negative impact of uncorrected accusations, with a weighted average unstandardized effect of 0.78 points (95% CI bounds = 0.41, 1.16), z = 4.07, p < .0001 (see Figure 4A), and no significant heterogeneity between studies, I^2 = 2.55%, Q = 4.10, p = .85. For *corrected accusations*, effects of inoculation on were not significant in full study 1 (b = 0.27, t = 0.72, p = .47), full study 2 (b = 0.38, t = 1.08, p = .28), or full study 3 (b = 0.20, t = 0.56, p = .57). Meta-analysis showed that the analogous effect for corrected accusations at a short delay trended in the same direction as the effect for uncorrected

accusations, but was not significant, with a weighted average unstandardized effect of 0.28 points (95% CI bounds = -0.08, 0.65), $z = 1.53$, $p = .13$ (see Figure 4B), and no heterogeneity between studies, $I^2 = 0.00\%$, $Q = 0.40$, $p = 1$.

An additional question of exploratory interest is how long-lasting the effects of inoculation are. In three of the experiments reported here (pilot studies 2 and 5, and full study 3, total n = 1572), participants provided a round of delayed ratings during a second session two days later. Examining only full study 3 (n = 1163), main effects were apparent for *uncorrected accusations* (b = -2.74, $t = -10.54$, $p < .0001$) and for *corrected accusations* (b = -0.91, $t = -3.49$, $p = .0005$), indicating that accusations were associated with lower ratings two days later regardless of whether they had been refuted, consistent with prior work. Similarly, a meta-analysis combining across all three experiments showed that candidates about whom *uncorrected accusations* were presented were rated lower than candidates with no accusation, with a weighted average unstandardized effect of -2.89 points (95% CI bounds = -5.72, -0.07), $z = -2.01$, $p = .045$, and significant heterogeneity between studies, $I^2 = 91.61\%$, $Q = 14.05$, $p = .0009$ (Supplemental Figure 3). Candidates about whom *corrected accusations* were presented were also rated lower than candidates with no accusation, with a weighted average unstandardized effect of -0.94 points (95% CI bounds = -1.39, -0.49), with no significant heterogeneity between studies, $I^2 = 0.0\%$, $Q = 0.08$, $p = .96$.

At the same time, there is no evidence to suggest inoculation interventions interact with the impact of accusations on ratings after a two-day delay (Supplemental Figure 4). Specifically, inoculation interventions did not reliably impact the effect of uncorrected accusations, either in full study 3 (b = 0.04, $t = 0.11$, $p = .92$), or in a meta-analysis, where the weighted average unstandardized effect is -0.02 points (95% confidence bounds = -0.70, 0.66), $z = -0.06$, $p = .95$. Inoculation also did not affect the response to corrected accusation candidates, either in full study 3 (b = -0.05, $t = -0.13$, $p = .90$), or in the meta-analysis, which showed a weighted average unstandardized effect of -0.01 points (95% confidence bounds = -0.69, 0.66), $z = -0.03$, $p = 0.97$.

Examination of Individual Differences

One factor that could hypothetically account for meaningful variability in the effectiveness of inoculation interventions is attention paid to the intervention. In other words, participants who paid close attention to the intervention may show greater impact of inoculation than those who view the videos more casually. Additionally, if samples differ on how likely participants were to pay close attention to the intervention, this factor could help to explain the observed heterogeneity between experiments. In all three full experiments and in pilot studies 7 and 8, close attention to the intervention was operationalized as accuracy on a difficult screening question about the control video to which all participants responded. Participants in pilot studies 4 and 5 were only shown one video (inoculation or control), so a difficult question corresponding to that video was used for the present analysis in those experiments. Preregistrations for full experiments 1 and 2

included a prediction that accuracy on a difficult screening question would moderate the impact of inoculation interventions.

The effect of inoculation on ratings of candidates given uncorrected accusations did interact with accuracy on the difficult question in full experiment 1 (b = 1.92, t = 2.22, p = .027), but not in full experiment 2 (b = 1.15, t = 1.39, p = .16), or full experiment 3 (b = 0.85, t = 1.02, p = .31). A meta-analysis that included all of these studies, as well as pilots 4, 5, 7, and 8, yielded a marginal effect, with a weighted average unstandardized effect of 1.62 points (95% confidence interval = -0.10, 3.34), z = 1.84, p = .065, and significant heterogeneity, I^2 = 59.16%, Q = 16.45, p = .012 (Supplemental Figure 5). If pilot 8 is excluded, due to having a notably smaller sample and being drawn from a different population, i.e., college students at a selective university, the meta-analytic effect is significant, with a weighted average unstandardized effect of 1.77 points (95% confidence interval = 0.56, 2.98), z = 2.86, p = .004, and no significant heterogeneity, I^2 = 34.32%, Q = 8.61, p = .13. The interaction between the effect of inoculation on ratings for corrected accusation candidates was not significant in full experiment 1 (b = 1.27, t = 1.47, p = .14), full experiment 2 (b = -0.64, t = -0.78, p = .43), or full experiment 3 (b = -0.86, t = -1.04, p = .30). A meta-analysis here across all 7 relevant experiments showed no effect, with a weighted average unstandardized effect of 0.35 points (95% confidence interval = -1.68, 2.37), z = 0.34, p = .74, and significant heterogeneity, I^2 = 68.87%, Q = 17.04, p = .009. Unlike for the uncorrected accusations, this effect did not change meaningfully when removing pilot study 8. Accuracy rates on the difficult screening question ranged from 45%-51% among the 6 experiments run online, with Pilot 8 having an accuracy rate of 58%. A chi-square test showed that accuracy rates among the different experiments did not differ from a random distribution (χ^2 = 7.20, df = 6, p = .30). Thus, while attentiveness to the intervention may predict how impactful the intervention was on candidates with uncorrected accusations, attentiveness to the intervention did not meaningfully vary between experiments, making it unlikely that this factor explains the heterogeneity in effect sizes between studies.

Another possible moderator variable is the political orientation of our participants. In full study 1, adding terms for interactions with political party identity yielded a significant 3-way interaction between this variable, inoculation condition, and corrected accusations (b = -0.58, t = -2.43, p = .015) indicating that inoculation was more effective at reducing the impact of corrected accusations in Democratic-leaning participants. The analogous interaction for uncorrected accusation stimuli was not significant (b = -0.21, t = -0.91, p = .37). However, the interaction between partisan identity and inoculation condition for corrected accusations did not replicate either in full study 2 (b = -0.01, t = -0.03, p = .98) or in full study 3 (b = 0.00, t = 0.00, p = 1). There was similarly no effect for uncorrected accusations in full study 2 (b = -0.09, t = -0.43, p = .67) or in full study 3 (b = 0.17, t = 0.91, p = 0.37). A meta-analysis that includes all 11 studies (Supplemental Figure 6) similarly does not indicate a reliable effect of partisanship on effects of inoculation on corrected accusations, with a weighted average unstandardized effect of -0.18 points (95% confidence interval = -0.44, 0.08), z = 1.37, p = .17. There was also no evidence of an

interaction in the analogous analysis for uncorrected accusations, with a weighted average unstandardized effect of -0.11 points (95% confidence interval: -0.83, 0.60), z = -0.31, p = 0.75. Thus, partisanship does not appear to consistently moderate the effectiveness of the intervention.

Headline Judgment Accuracy

While our primary interest was in whether inoculation would affect attitudes towards hypothetical political candidates, we also examined whether the interventions impacted scores on the MIST-20 headline judgment accuracy measure. This measure examines people's ability to distinguish prototypically false headlines from true headlines. The effect of inoculation condition on MIST accuracy scores (combining across true and false items) was not significant in individual large experiments: full study 1 (t(1448) = 0.376, p = .71), full study 2 (t(1531) = 0.141, p = .88), or full study 3 (t(1161) = 1.64, p = .10). However, a meta-analysis (Figure 5A) showed a small but significant effect on this measure when combining across all 11 studies, with a weighted average mean difference of 0.86% accuracy (95% confidence interval: 0.03%, 1.69%), z = 2.03, p = .043, and no significant heterogeneity between studies, I^2 = 0.0%, Q = 7.24, p = .70. This effect appears to be driven primarily by improvements in recognizing false items, as a meta-analysis across all 11 studies (Figure 5B) finds that these items show a significant increase in accuracy of 1.32% (95% confidence interval: 0.26%, 2.38%), z = 2.45, p = .014, with no significant heterogeneity between studies, I^2 = 0.27%, Q = 10.85, p = .37. No significant change in accuracy was apparent for real items (Figure 5C); the weighted mean effect was an accuracy increase of 0.57% after inoculation (95% confidence interval: -0.46%, 1.60%), z = 1.08, p = .28, with no significant heterogeneity between studies, I^2 = 0.00%, Q = 2.81, p = .99. These effects should be interpreted cautiously, however, because in full study 3, as well as in pilot 2, the MIST was administered in the second session two days later, and we cannot rule out the possibility of bias in who chose to return for the second session in the inoculation group versus the control group. When these two experiments are removed, the meta-analytic effects are no longer significant, either for all items or for false items.

Memory Test

A memory test, completed two days after the initial experimental session, was included in full study 3. This was of interest based on the hypothesis that the memory advantage in recognizing accusations would reflect emotional activation in response to the stimulus, which hypothetically would be reduced by inoculation. These data were first analyzed using a 2 (Impactfulness: Accusation/Refutation vs. Control) × 2 (Post 1 vs. Post 2) x 2 (Inoculation condition) mixed ANOVA, with repeated measures on the first two factors. Across all participants, the expected Impactfulness x Post interaction was significant (F(1, 1161) = 50.07, p < .001, η_p^2 = .041), showing that the degree to which accusation stimuli are remembered better than post 1 control stimuli is greater than the degree to which refutation stimuli are remembered better than post 2 control stimuli (Figure 6). This replicates the basic memory effect reported previously by Cohen et al. (2024). The 3-way interaction (Impactfulness x Post x Inoculation condition) was also significant (F(1, 1161) = 7.37, p = .007, η_p^2 = .006). However, while it was predicted that the

Impactfulness x Post interaction would be weaker after inoculation, due to inoculation leading to less emotional activation in response to accusation stimuli, the effect was actually stronger in the inoculation condition, contrary to our predictions. The two-way Impactfulness x Post interaction was significant both when no inoculation video was shown ($F(1, 633) = 9.96$, $p = .002$, $\eta_p^2 = .015$) and also for participants in the inoculation condition ($F(1, 528) = 46.73$, $p < .001$, $\eta_p^2 = .081$). Further analysis using separate 2 x 2 mixed ANOVAs for post 1 and post 2 showed that the magnitude of memory benefit for accusations did not differ based on inoculation condition ($F(1, 1161) < 1$, $p = .67$, $\eta_p^2 = .000$), but the memory benefit for refutations did interact with inoculation condition ($F(1, 1161) = 12.49$, $p < .001$, $\eta_p^2 = .011$), with the control condition showing a greater memory benefit. We had hypothesized that inoculation might increase the memorability of refutations by increasing their importance, but again this effect went in the opposite direction as predicted. Finally, it was predicted that the memory benefit for accusation stimuli relative to control post 1 stimuli would correlate with the magnitude of Continued Influence Effects on immediate ratings (i.e., the difference in ratings between corrected accusation and no accusation candidates), but there was no evidence here to support such an effect ($r = -0.003$, $p = .91$). Thus, while accusations were remembered better than refutations, differences in the magnitude of this memory effect by inoculation condition or by individual were not able to provide insight into mechanistic underpinnings of inoculation interventions or CIEs.

Discussion

Inoculation treatments that promote critical thinking about common manipulation tactics appear to reduce candidates' vulnerability to unproven accusations. These interventions clearly attenuate the impact of accusations on candidate impressions when no correction has been provided, i.e., when the veracity of the accusation remains ambiguous. There also appears to be a small but significant positive effect of inoculation on the impact of *corrected* accusations on candidate impressions. This latter finding suggests that inoculation interventions either operate via a different mechanism from factual corrections in promoting skepticism about the accusation, or enhance the weight individuals assign to corrections when forming impression of candidates.

The attenuating effect of inoculation interventions on the impact of uncorrected accusations remains evident when candidate ratings are collected after a short delay. It should be noted that in most studies that included delayed ratings (all 3 full experiments and pilot studies 4, 5, 7, and 8, but *not* pilot studies 3 or 6), participants were given a brief "booster" inoculation prior to the delayed ratings. This design was intended to provide a best-case scenario of whether inoculation effects could persist beyond the immediate judgment. Still, it does leave some ambiguity in the interpretation. One possibility is that the inoculation leads to a reduction in the initial emotional impact of accusations, which would be maintained even without a "booster". Alternatively, participants may remember that a particular candidate was targeted by an uncorrected accusation, but the booster is necessary to remind participants to reduce the impact of emotional content on their ratings. Effects on delayed ratings were also relatively small in magnitude and somewhat

inconsistent across studies even in this best-case scenario. The analogous effect for candidates targeted by corrected accusations was not statistically significant even after combining across studies. Furthermore, effects of inoculation are not apparent in either condition after a two-day delay, though this conclusion should be made tentatively because this effect was only examined in a subset of experiments. Taken together, these findings suggest that while inoculation benefits can persist after a short delay on the order of minutes, their longer-term durability in reducing the impact of accusations on attitudes remains uncertain (cf., [27]-[28]).

This work provides the first evidence that inoculation interventions can reliably reduce the extent to which unsubstantiated accusations against political candidates undermine subjective evaluations of those candidates. The theoretical framework proposed in [18] helps to situate these results within a broader account of belief formation and change. Within this framework, inoculation can be understood as first weakening belief in the accusations, which in turn reduces the downstream impact of those beliefs on attitudes. Given the close alignment between attitudes and behavioral intentions found in the prior work with this paradigm [2], we can assume that when inoculation reduces the impact of unsubstantiated accusations on attitudes, it also reduces their impact on behavioral intentions towards targeted political candidates. In other words, by establishing the effects of inoculation on attitudes, the present findings represent an important step toward demonstrating corresponding effects on political behavior.

We also provide novel evidence on how inoculation affects headline judgment accuracy for stimuli that were not intended to exemplify specific persuasive techniques targeted by the intervention. Despite work firmly showing that inoculation improves headline judgment accuracy more broadly, it is not established whether it can improve performance on the MIST. One study found that on an 8-item version of the MIST, an inoculation game improved accuracy for false headlines while reducing accuracy for real headlines, suggesting a change in criterion rather than an improvement in discernment [21]. At the same time, other work has suggested that inoculation can improve MIST-8 accuracy after feedback has been provided [29]. In the present study, inoculation improved accuracy for false headlines without reducing accuracy for true headlines, leading to a small but significant increase in overall accuracy. Still, the effect size was extremely small, and this effect may in part reflect a selection bias in the participants who returned for the delayed session, which, in some experiments, is when the MIST was administered. These findings suggest that transfer of inoculation effects to novel stimuli is generally modest, rather than effects of inoculation only being small with respect to attitudes.

A further goal of this work was to investigate the mechanisms underlying how inoculation interventions work and why Continued Influence Effects occur in this context. We replicated prior findings using the same paradigm that accusation stimuli are remembered better than refutation stimuli, relative to matched control content, particularly when memory is assessed after a two-day delay [2]. We interpret this observed

difference as resulting from accusations producing a greater emotional response than refutations, which strengthens memory encoding and consolidation.

There was further tentative evidence in the prior work suggesting that the magnitude of the memory effect might correlate with the magnitude of Continued Influence Effects, and thus, that memory effects could serve as a proxy for understanding the mechanisms of inoculation interventions. This hypothesis was not supported here, in part because the previously observed association between the memory impact of accusations and their attitudinal impact did not replicate. The underlying mechanisms by which inoculation reduces the impact of unproven accusations on attitudes toward hypothetical political candidates warrant further investigation.

It is also important to note that, for immediate ratings, there was substantial heterogeneity in effect sizes across studies, both in the baseline condition and in estimates of the inoculation intervention's effects for both uncorrected and corrected accusations. The sources of this heterogeneity are unclear. Although there were minor procedural differences across studies, these variations were subtle, and no systematic relationship between effect size and identifiable study characteristics emerged that could be formally tested as a moderator. Additionally, two factors that could hypothetically moderate effects of inoculation interventions, attention to the intervention and partisan political orientation, did not consistently do so. Thus, we chose to assume that combining across all studies, despite the heterogeneity, was the best of imperfect options to determine the true effect size.

One other potential complication is that the strongest effects of inoculation were observed on candidates whose accusations were never refuted, i.e., those with uncorrected accusations. It is arguable whether such uncorrected accusations should be perceived as true or false. From the perspective of some misinformation researchers, that the primary goal of interventions is to help people discern clearly true from clearly false information (e.g., [30]), reducing the impact of uncorrected accusations may not be seen as beneficial. However, others argue that misleading information, even when originating from a high-quality news source, can cause greater overall harm than overtly false information [31]. It is also important to recognize that not all false accusations can be thoroughly fact checked, and even when they are, not all voters will be exposed to or believe those fact checks. Furthermore, evidence suggests that false information tends to employ stronger negative emotional and moral appeals than true information [32]-[33]. The accusation stimuli used in the present studies included these features, regardless of whether the accusation is explicitly refuted.

An intervention that reduces the weight of emotionally charged content on decision-making is likely to be a beneficial to the information ecosystem, even when a dramatic accusation ultimately turns out to be true. Inoculation interventions do not fully ameliorate the impact of uncorrected accusations on candidate ratings, nor should they. Indeed, if evidence is later revealed to support an accusation, it is rational for voters to penalize the

candidate. Inoculation does not prevent this outcome, rather, it slows the process, allowing journalists the opportunity to confirm or refute a claim before a candidate's support entirely collapses due to an unverified accusation.

References

1. Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33,* 460-480.
2. Cohen, M. S., Halewicz, V., Yildirim, E., & Kable, J. W. (2024). Continued influence of false accusations in forming impressions of political candidates. *PNAS Nexus*, *3*(11), pgae490.
3. Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*, 348–384.
4. Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological Inoculation against Misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science, 700*, 136-151.
5. McGuire, W. J. (1964). Some contemporary approaches. In: *Advances in experimental social psychology* (Vol. 1, pp. 191-229). Academic Press.
6. Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE, 12,* e0175799.
7. Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Nature Humanities and Social Sciences Communications, 5,* 65.
8. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8,* eabo6254.
9. Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data and Society, 8*(1).
10. Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General, 152(9),* 2411–2437.
11. Simchon, A., Zipori, T., Teitelbaum, L., Lewandowsky, S., & van der Linden, S. (2026). A signal detection theory meta-analysis of psychological inoculation against misinformation. *Current Opinion in Psychology, 67,* 102194.
12. Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20,* 1420–1436.
13. Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., Cook, J. (2012). Misinformation and its correction—Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13,* 106–131.
14. Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. (2022). The psychological drivers of

misinformation belief and its resistance to correction. *Nature Reviews Psychology,* 1, 13-29.
15. van der Linden, S. L., Leiserowitz, A. A., Rosenthal, S. A., Feinberg, G. D., & Maibach, E. W. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1, 1600008.
16. Cook, J., Lepage, C., Hopkins, K. L., Cook, W., Kolog, E. A., Thomson, A., Iddrisu, I., & Burnette, S. (2024). Co-designing and pilot testing a digital game to improve vaccine attitudes and misinformation resistance in Ghana. *Human Vaccines & Immunotherapeutics, 20,* 2407204.
17. Spampatti, T., Hahnel, U.J.J., Trutnevyte, E., & Brasch, T. (2024). Psychological inoculation strategies to fight climate disinformation across 12 countries. *Nature Human Behavior*, 8, 380–398 (2024).
18. Granados Samayoa, J.A. & Albarracín, D. (2025). Understanding belief-behavior correspondence: Introducing a belief-to-behavior process model. *Psychological Inquiry, 36(1),* 1-22.
19. Hamilton, K., & Hagger, M. S. (2025). Elucidating the processes by which specific beliefs and attitudes predict behavior: Advocacy for a comprehensive, integrated approach. *Psychological Inquiry, 36(1),* 26-35.
20. Granados Samayoa, J.A. & Albarracín, D. (2025). Beliefs and belief-to-behavior inferences: Clarifications, rebuttals, and extensions. *Psychological Inquiry, 36(1),* 75-80.
21. Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2024). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods, 56,* 1863-1899.
22. Garrett, R. K. & Weeks, B. E. (2017). Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLoS ONE, 12,* e0184733.
23. Newton, C., Feeney, J., & Pennycook, G. (2024). On the disposition to think analytically: four distinct intuitive-analytic thinking styles. *Personality and Social Psychology Bulletin*, *50*, 906-923.
24. Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry, 84,* 443-451.
25. Duckitt, J., Wagner, C., du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology, 83*, 75–93.
26. Lauriola, M., Foschi, R., Mosca, O., & Weller, J. (2016). Attitude toward ambiguity: Empirically robust factors in self-report personality scales. *Assessment, 23,* 353-373.
27. Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied, 27(1),* 1-16.
28. Maertens, R., Roozenbeek, J., Simons, J.S., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2025). Psychological booster shots targeting memory

increase long-term resistance against misinformation. *Nature Communications, 16,* 2062.
29. Leder, J., Schellinger, L. V., Maertens, R., van der Linden, S., Chryst, B., & Roozenbeek, J. (2024). Feedback exercises boost discernment of misinformation for gamified inoculation interventions.*Journal of Experimental Psychology: General, 153*(8), 2068–2087.
30. Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behavior, 7,* 1231-1233.
31. van der Linden, S., and Kyrychenko, Y. (2024). A broader view of misinformation reveals potential for intervention. *Science*, 384, 959-960.
32. Carrasco-Farré, C. (2022) The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Science Communications, 9,* 162.
33. McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991–996.

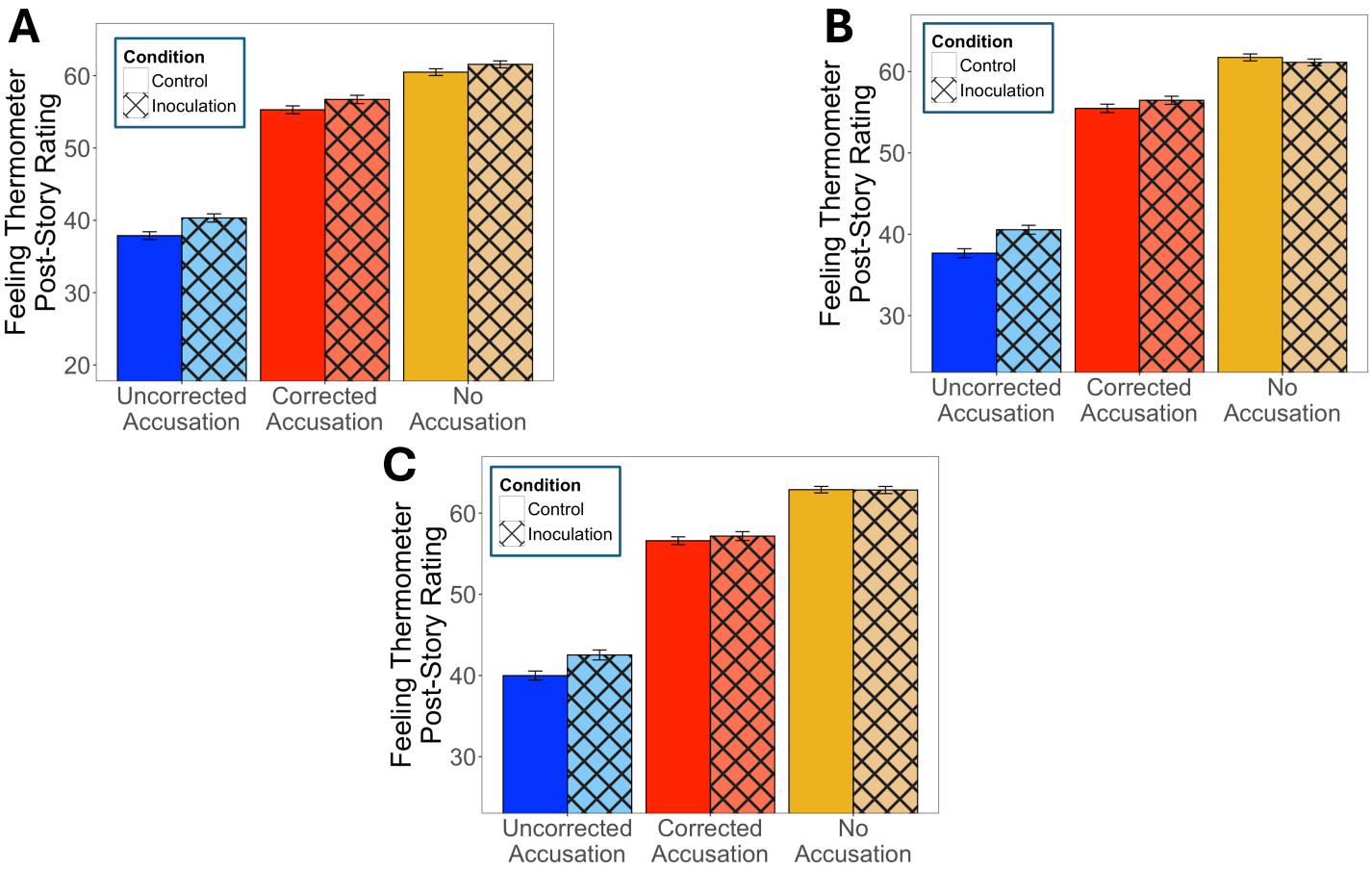


Figure 1. Mean immediate feeling thermometer ratings by condition in (A) Full Study 1, (B) Full Study 2, and (C) Full Study 3. All three experiments show a steep decline in ratings after an uncorrected accusation, and a smaller but still notable decline in ratings after a corrected accusation, relative to the no accusation condition. Inoculation consistently lessens the decline in ratings for candidates facing uncorrected accusations, relative to candidates in the no accusation condition. Effects on candidates facing corrected accusations are more variable; see Figure 2B.

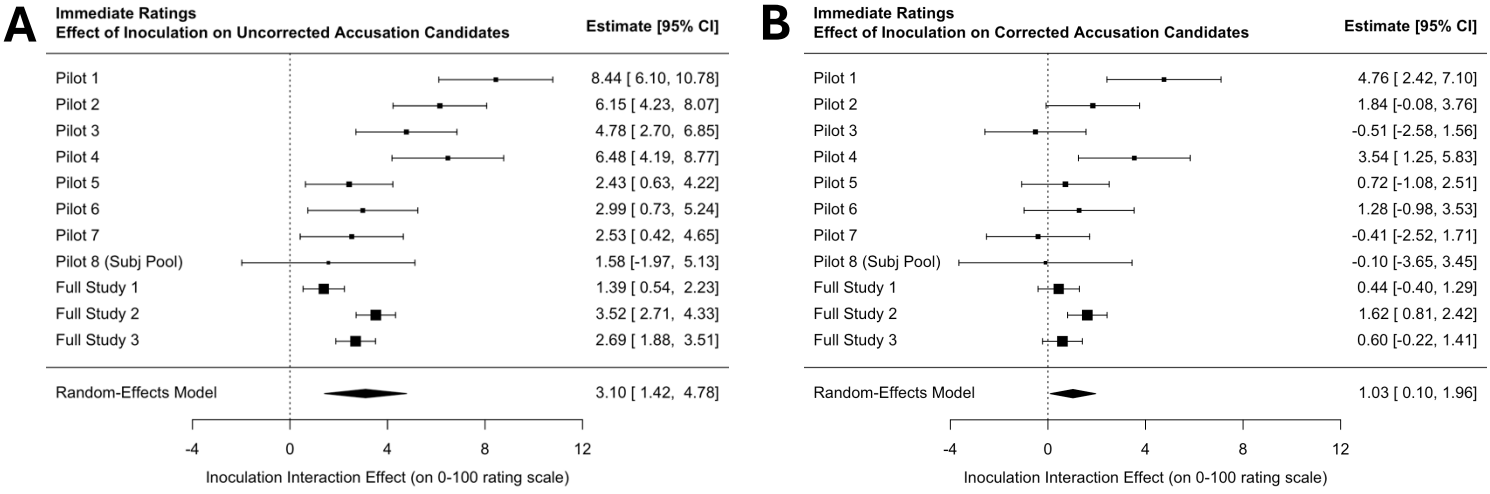


Figure 2. Interaction effects between treatment effect of inoculation and (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, are both significant for immediate candidate ratings.

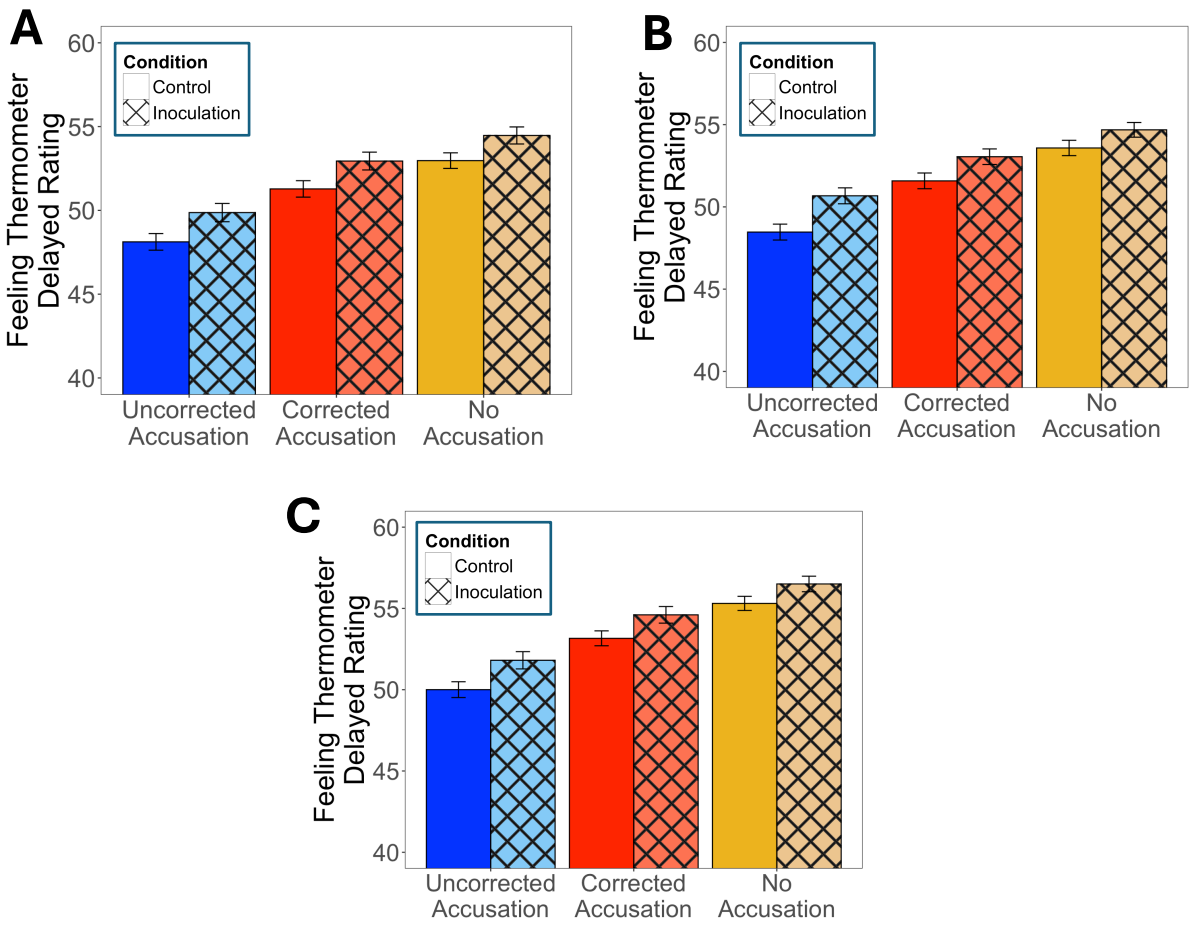


Figure 3. Mean short-delay feeling thermometer ratings by condition in (A) Full Study 1, (B) Full Study 2, and (C) Full Study 3. All three experiments continue to show a decline in ratings after both uncorrected and corrected accusations, relative to the no accusation condition. Effects of inoculation are smaller but are still evident at a short delay.

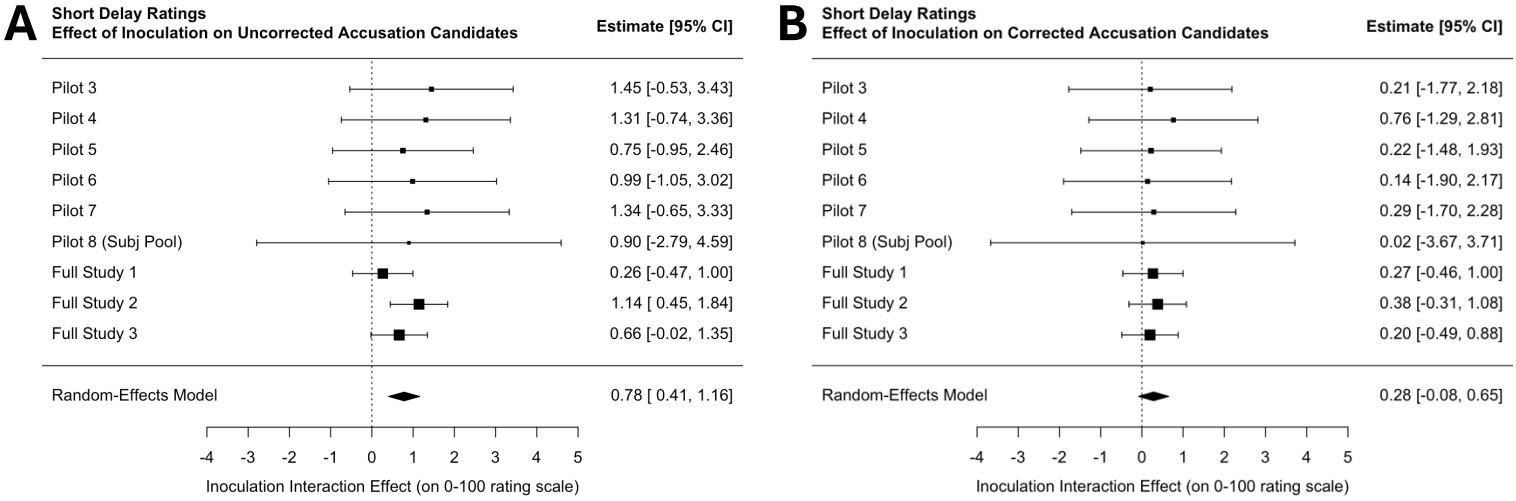


Figure 4. Combining all data in a meta-analysis, interaction effects between treatment effect of inoculation and Uncorrected Accusation condition (A) are significant on short-delay candidate ratings, while candidates in the Corrected Accusation condition (B) show a trend but no significant effect.

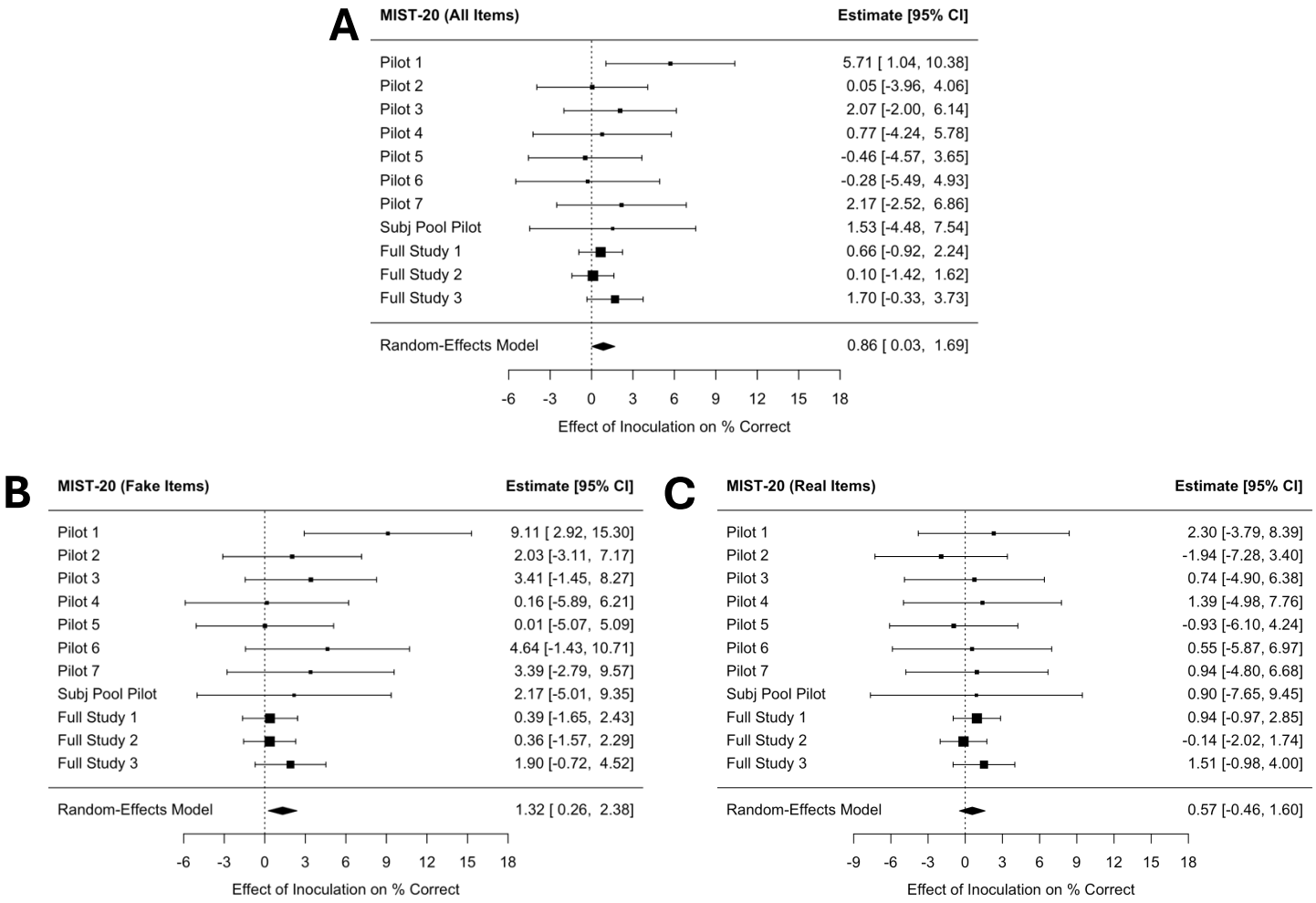


Figure 5. Effects of inoculation on MIST-20 headline judgment accuracy measure. (A) Inoculation led to a small but significant increase in overall accuracy. This effect was driven primarily by improved accuracy for fake items after inoculation (B), as real items showed no effect of inoculation on accuracy (C).

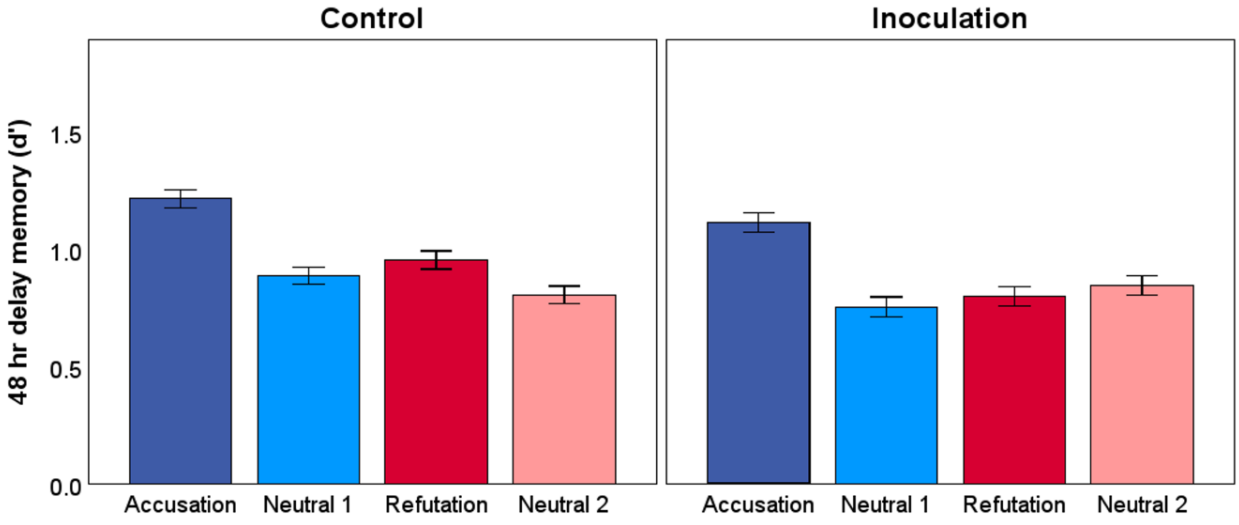
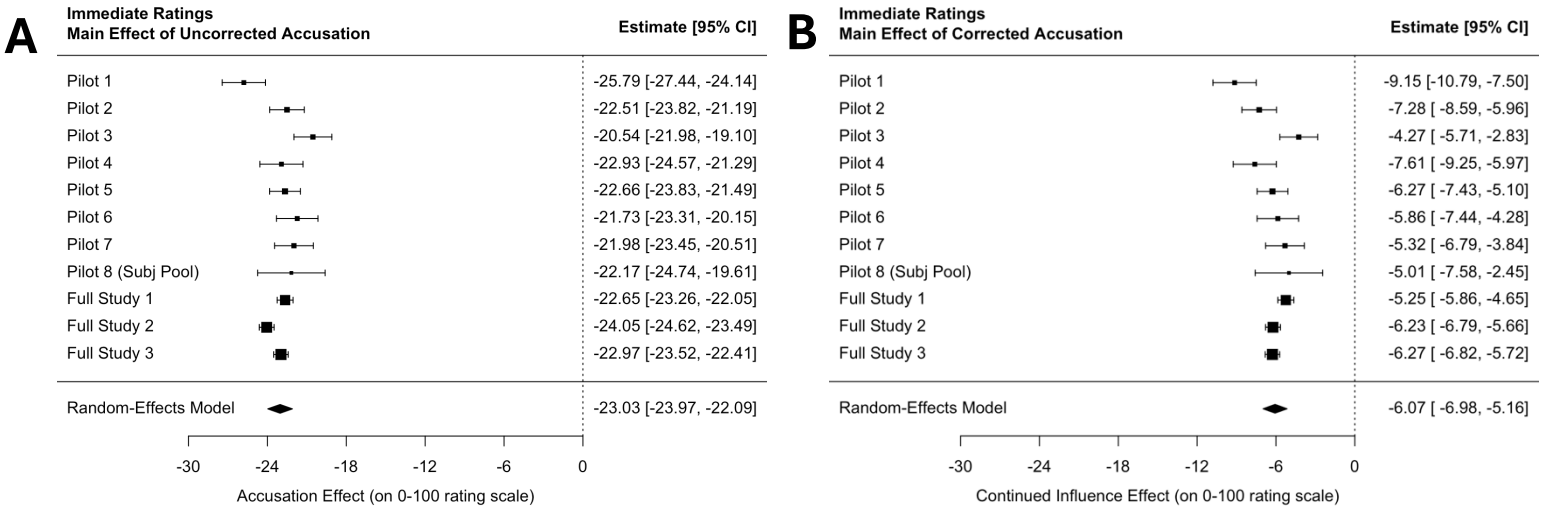
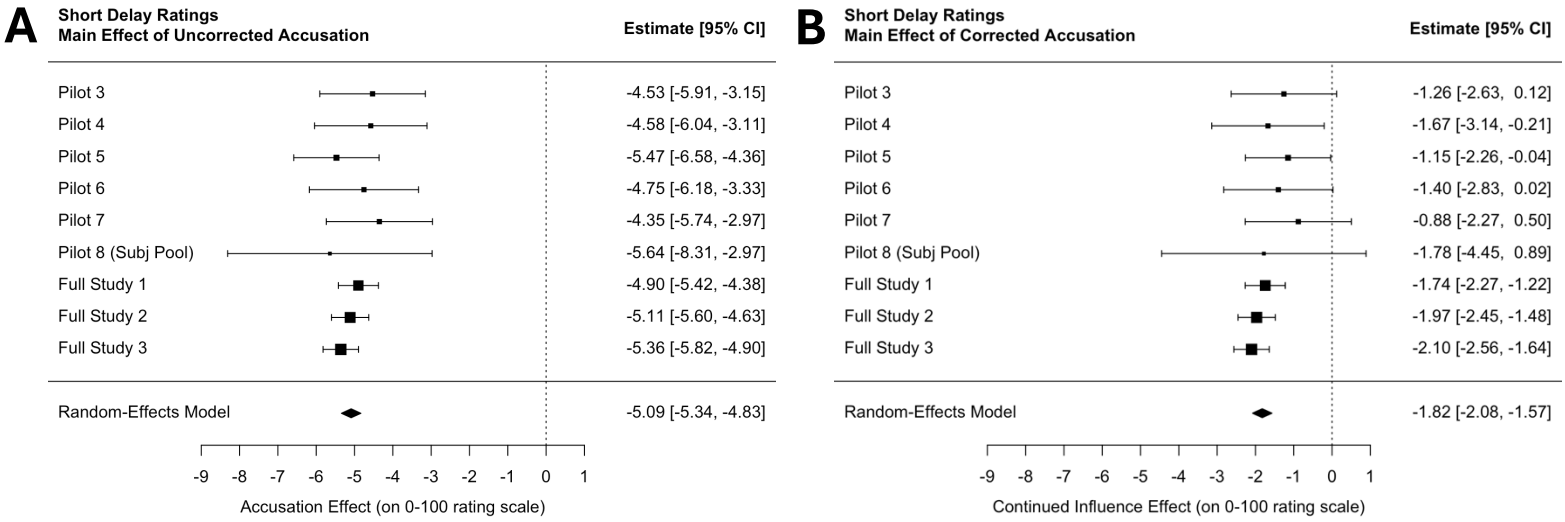


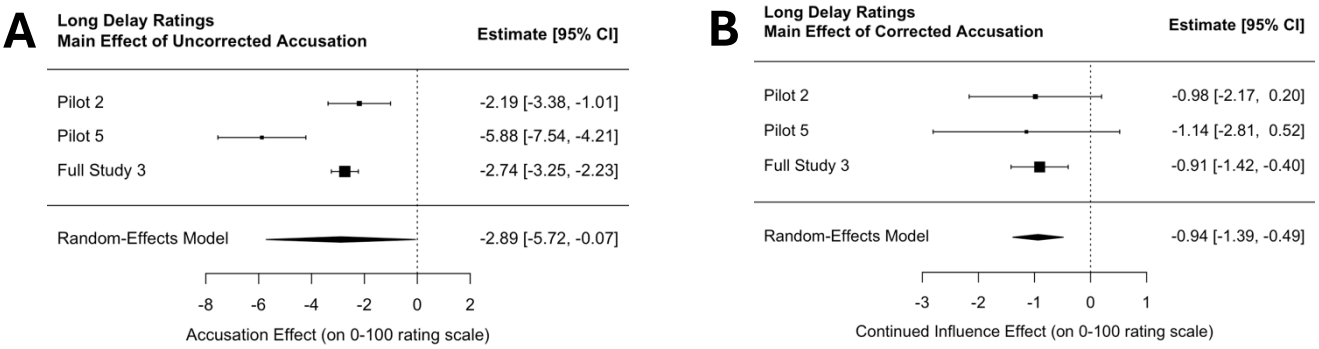
Figure 6. Recognition memory for each type of stimulus, measured after a ~ 48 hour delay in Full Study 3. In both conditions, accusation stimuli were remembered better than matched control stimuli, and this effect was larger than the corresponding effect for refutation stimuli.



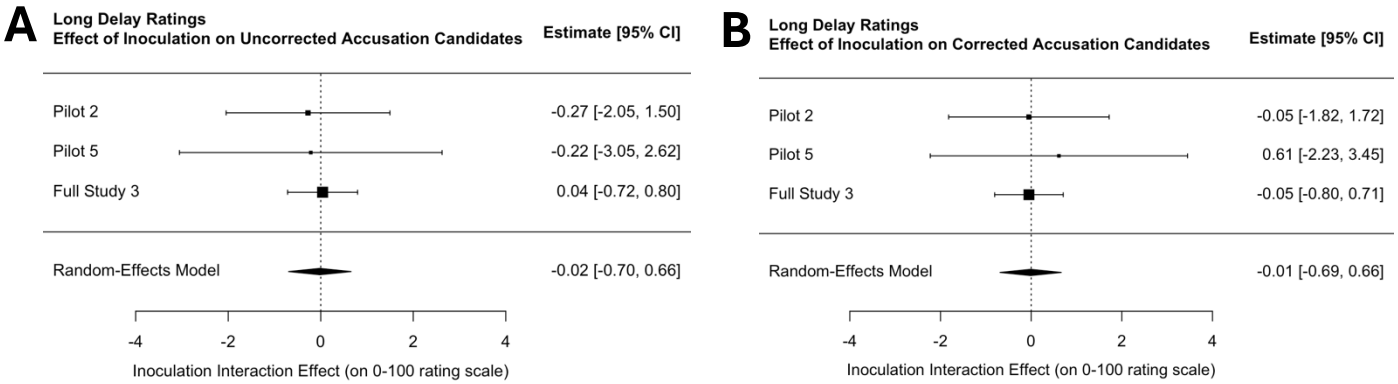
Supplemental Figure 1. Main effects of (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on immediate candidate ratings, relative to candidates with No Accusation, in the baseline (no-inoculation) condition



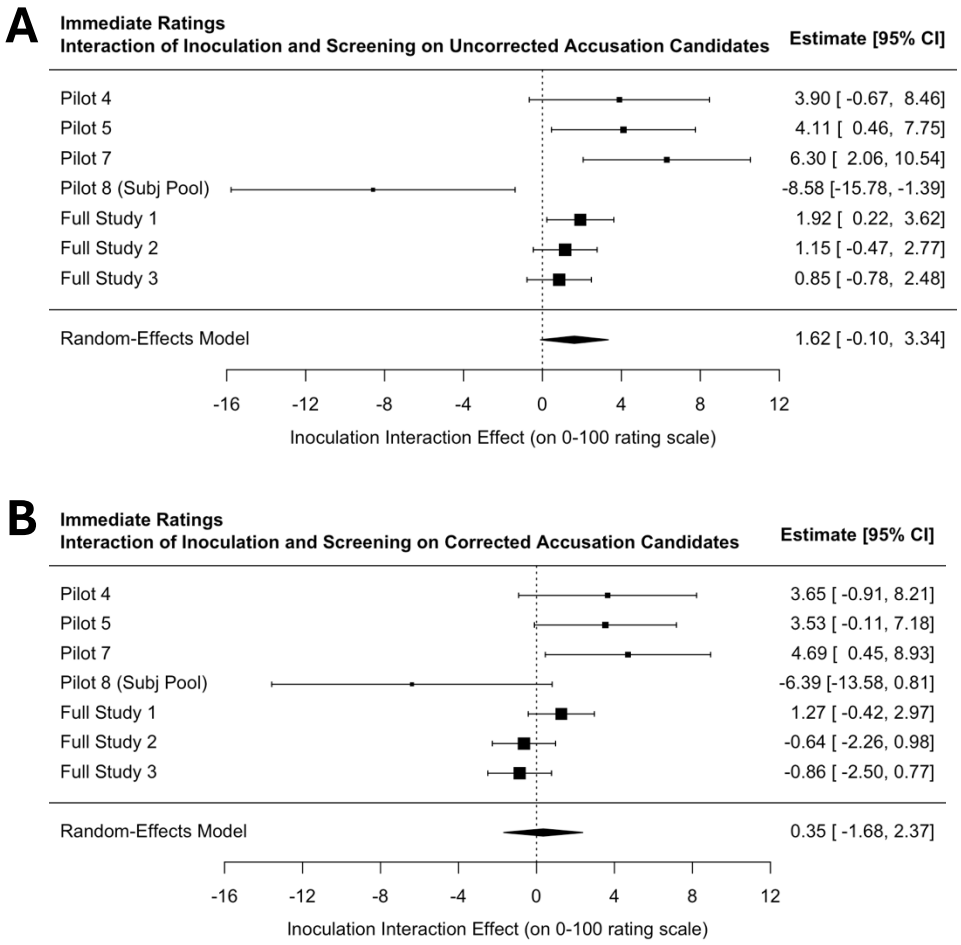
Supplemental Figure 2. Main effects of (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on short-delay candidate ratings, relative to candidates with No Accusation, in the baseline (no-inoculation) condition.



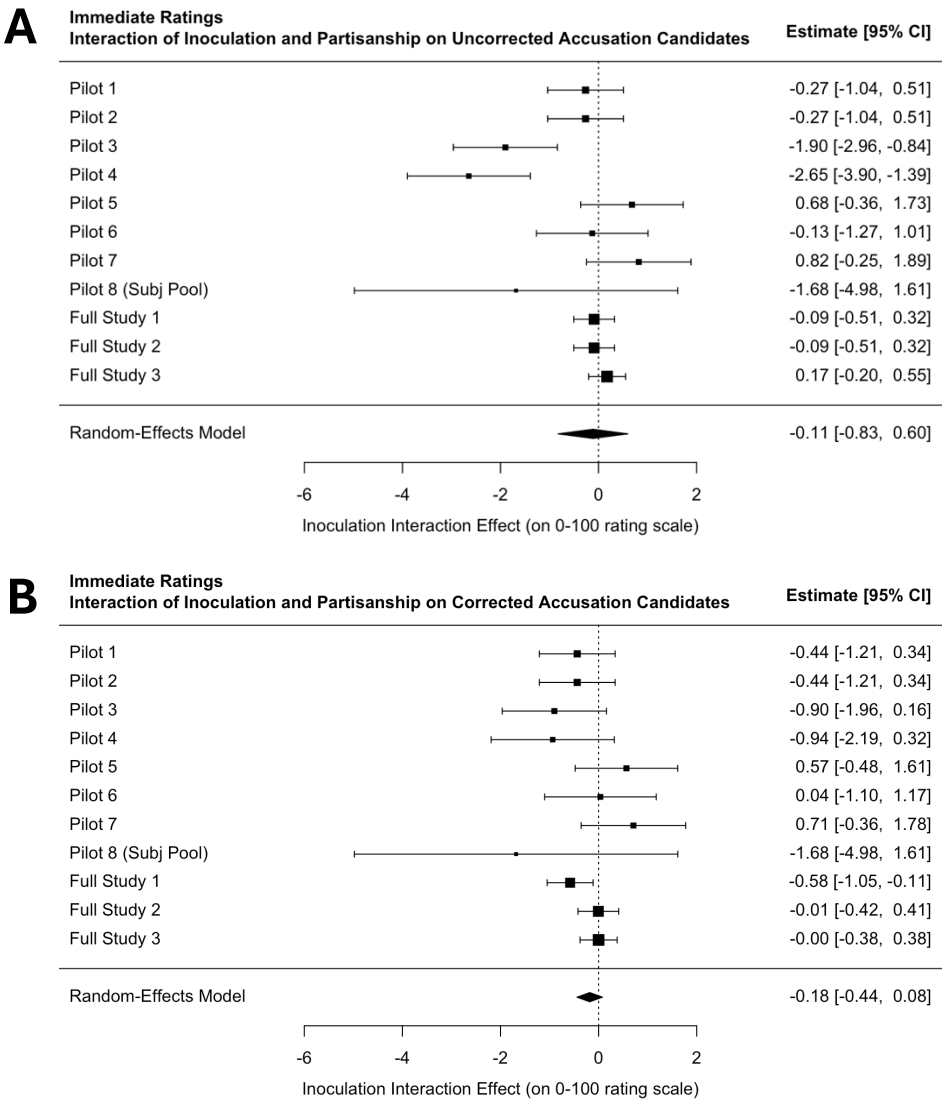
Supplemental Figure 3. Main effects of (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on long-delay candidate ratings, relative to candidates with No Accusation, in the baseline (no-inoculation) condition.



Supplemental Figure 4. Interaction effects between treatment effect of inoculation and (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on long-delay candidate ratings.



Supplemental Figure 5. Three-way interaction effects between accuracy on difficult screening questions, treatment effect of inoculation, and (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on immediate candidate ratings. Note that when data from Pilot 8 is excluded, due to its clear outlier status and its sample being drawn from a participant pool at a selective university rather than an open online participant pool, the meta-analytic effect shown in panel A is statistically significant.



Supplemental Figure 6. Three-way interaction effects between political partisanship, treatment effect of inoculation, and (A) Uncorrected Accusation condition, and (B) Corrected Accusation condition, on immediate candidate ratings.

Supplemental Table 1. Participant recruitment information.

Study	Sample size	Compensation	Source
Pilot 1	n = 201 total, n = 200 final	$10.00	CloudResearch Connect
Pilot 2	n = 310 total, n = 310 final (Part 1)	$9.00 (Part 1) + $5.00 (Part 2)	CloudResearch Connect
Pilot 3	n = 244 total, n = 244 final	$12.00	Prolific
Pilot 4	n = 200 total, n = 199 final	$8.00	CloudResearch Connect
Pilot 5	n = 298 total, n = 296 final (Part 1)	$9.00 (Part 1) + $5.00 (Part 2)	CloudResearch Connect
Pilot 6	n = 200 total, n = 198 final	$9.00	CloudResearch Connect
Pilot 7	n = 241 total, n = 227 final	$9.00	CloudResearch Connect
Pilot 8	n = 70 total, n = 67 final	No cash compensation	UChicago Subject Pool
Full Study 1	n = 1501 total, n = 1447 final	$5.00	CloudResearch Connect
Full Study 2	n = 1602 total, n = 1528 final	$5.00	CloudResearch Connect
Full Study 3	n = 1605 total, n = 1539 final (Part 1)	$5.50 (Part 1) + $4.50 (Part 2)	CloudResearch Connect

Supplemental Table 2. Procedural details by experiment

	Inoculation procedure	Before inoculation	After inoculation, posts, & ratings
Pilot 1	*Bad News* game (inoculation) or Tetris (control), varied between-subjects	(none)	Sharing judgments, epistemic beliefs, CTSQ, digital literacy, affective polarization, MIST-20, metacognitive dots task
Pilot 2	*Bad News* game (inoculation) or Tetris (control), varied between-subjects	(none)	**In session 1:** Sharing judgments, epistemic beliefs, CTSQ, digital literacy, affective polarization, metacognitive dots task **In session 2:** Choice task, delayed ratings, MIST-20
Pilot 3	*Bad News* game (inoculation) or Tetris (control), varied between-subjects	(none)	Epistemic beliefs, CTSQ, digital literacy, affective polarization, MIST-20, choice task, delayed ratings
Pilot 4	Inoculation video vs. control video, varied between subjects, followed by difficult question on the viewed video	(none)	Epistemic beliefs, CTSQ, digital literacy, affective polarization, MIST-20, inoculation booster, choice task, delayed ratings
Pilot 5	Inoculation video vs. control video, varied between subjects, followed by difficult question on the viewed video	(none)	**In session 1:** Epistemic beliefs, dangerous world beliefs, moral absolutism, affective polarization, digital literacy, sharing judgments, MIST-20, inoculation booster, choice task, delayed ratings **In session 2:** Choice task, delayed ratings
Pilot 6	Inoculation video vs. control video, varied between subjects, followed by two easy questions on the viewed video	(none)	Epistemic beliefs, dangerous world beliefs, moral absolutism, affective polarization, digital literacy, sharing judgments, MIST-20, choice task, delayed ratings
Pilot 7	All participants saw control video, answered both 1 difficult and 1 easy question about it; Participants in inoculation condition then saw inoculation video, and answered 1 difficult and 1 easy question about it	Epistemic beliefs, dangerous world beliefs, digital literacy, moral absolutism, affective polarization	Inoculation booster, sharing judgments, MIST-20, choice task, delayed ratings

Pilot 8	(same as above)	Epistemic beliefs, digital literacy	Inoculation booster, MIST-20, choice task, delayed ratings
Full Study 1	(same as above)	Digital literacy	Inoculation booster, MIST-20, delayed ratings
Full Study 2	(same as above)	(none)	Inoculation booster, MIST-20, delayed ratings
Full Study 3	(same as above)	(none)	**In session 1:** Inoculation booster, delayed ratings **In session 2:** Delayed ratings, memory test, MIST-20