

# The Sophistication Paradox: Political Sophistication Amplifies, Rather Than Buffers, Polarization-Driven Misinformation Belief

Atif Ur Rehman

atifurrehman@hust.edu.cn

Huazhong University of Science and Technology

---

## Research Article

**Keywords:** Affective Polarization, Misinformation Belief, Political Sophistication, Motivated Reasoning, Partisan Asymmetry, Fractional Logit Regression, Equivalence Testing

**Posted Date:** October 14th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10939955/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Political misinformation and affective polarization are considered entangled threats to democratic governance; however, evidence linking the two has relied on single-sample designs, measurement approaches poorly suited to bounded belief outcomes, and correlative analyses that preclude causal inference. The present study examined the extent to which affective polarization predicts belief in political misinformation; the extent to which this relationship may differ for in-party versus out-party content; the extent to which political sophistication acts as a buffer or intensifier of the relationship; and whether the relationship suggests a causal pathway that can be manipulated. Three independent data sets were examined: two national observational samples (Social Media  $N = 5,104$ ; Time Series/ANES 2020,  $N = 6,352$ ) and one two-arm randomized experiment ( $N = 840$ ). Results revealed that affective polarization was correlated with increased belief in in-party misinformation and reduced belief in out-party misinformation; both results were significant and of the same sign and magnitude across both observational samples (all  $p < .001$ ). Results were robust across families and model specifications. The causal structure of this association suggests that it was not a matter of affect but of stable partisan identity, with implications for the design and targeting of misinformation-correction interventions.

## 1. Introduction

### 1.1 Background and Problem Statement

Political misinformation and affective polarization have received considerable attention as an unprecedented twin challenge to democratic life in the modern era, yet they have mostly been examined separately. In the United States, affective polarization — a tendency to distrust, dislike, and morally condemn members of the opposition party — has steadily increased over the past 20 years and is now comparable to, or greater than, the racially based animus generally associated with racial polarization (Iyengar et al., 2019; Iyengar & Westwood, 2015). Ideological polarization is a matter of disagreement, whereas affective polarization appears to be essentially an identity phenomenon; partisanship generates in-group and out-group feelings, irrespective of ideology (Iyengar et al., 2012). This hostility is not limited to political judgments, especially when it stems from identity. It influences hiring and dating choices, opposition to public health counsel, and other decisions, and it has been experimentally shown to be as strong as, or stronger than, polarization on race (Iyengar & Westwood, 2015).

There has been a similar scholarly and public consensus of alarm about misinformation. False and unsubstantiated claims spread in homogeneous and polarized communities that serve as echo chambers, as shown in large-scale network analyses of Facebook and Twitter data, where both homophily and polarization are correlated with the size of misinformation cascades (Del Vicario et al., 2016). Field-experimental studies reveal that partisan exposure to cross-cutting political information does not uniformly reduce polarization and can, in fact, exacerbate it (Bail et al., 2018). Large observational analyses of Facebook news feed data indicate that algorithmic curation (Bakshy et al., 2015) and, especially, individual choice reduce exposure to ideologically diverse information. At the

individual level, misinformation seems to spread faster than accurate information online, partly because of the novelty and emotional appeal of false claims (Vosoughi et al., 2018). Altogether, this literature highlights that affective polarization and belief in misinformation are two mutually reinforcing elements of the current information landscape, rather than distinct pathologies (Finkel et al., 2020).

What remains far less settled is the precise causal and moderating architecture linking the two constructs. Does affective polarization cause susceptibility to misinformation, or does the relationship stem from an underlying factor, such as the strength of partisan identity? Is the relationship similar when it comes to misinformation that favors the party and when it comes to misinformation that hurts the other party, or is the repression mechanism a partisan, internal asymmetry? Importantly, does political sophistication—precisely the kind of sophistication that is thought to buffer citizens against false belief, also in this case—provide the conceptual resources necessary to rationalize identity-congruent falsehoods (Kahan et al., 2017)?

## 1.2 Limitations of Existing Research

Although the literature is well established, there are multiple caveats to studying affective polarization as directly related to beliefs about misinformation. Most prior research is conducted as a single-sample study, which complicates the interpretation of associations, as they may represent an overall population phenomenon or merely particular characteristics of a single dataset, platform, or election cycle (Guess et al., 2019; Osmundsen et al., 2021). The standard for establishing external validity is cross-dataset replication with the same model specifications, and this standard is rarely met in this literature.

Furthermore, misinformation-belief indices are nearly always continuous proportions on the [0,1] interval, and ordinary least squares (OLS) remains the literature's default estimator. The OLS fit to a bounded outcome yields out-of-range predicted probabilities, heteroskedastic and biased SEs at the extremes of its range, and systematically misleading inference in the regime where polarization is highest and belief in misinformation is greatest; the high-polarization/belief-in-misinformation regime is thus precisely the one that is problematic substantively. (Papke & Wooldridge, 1996; Smithson & Verkuilen, 2006).

The empirical results relating to the moderating effect of political sophistication remain unsettled, and the theoretical explanations are debatable. A traditional view holds that sophistication, numeracy, and education serve as vaccines against misinformation, with the notion that as more is known, greater cognitive ability leads to better assessment of evidence (Pennycook & Rand, 2021). However, a competing (and perhaps more algorithmic) view, rooted in motivated cognition theory, suggests that sophistication, rather, affords more sophisticated means of motivated cognition, and thus that those with the most knowledge are the most polarized, not the least (Kahan et al., 2017; Taber & Lodge, 2006). This moderation has not been previously examined with the formal interaction term based on conditional marginal effects across independent samples.

The causal status of the polarization–misinformation link has rarely been evaluated experimentally. Nyhan and Reifler (2010) demonstrated that fact-based corrections frequently fail to reduce partisan

misperceptions and can even backfire, but this and related designs manipulate the correction rather than the underlying affective state, leaving the reverse causal question—whether manipulating partisan affect itself shifts misinformation belief—largely untested. Relatedly, experiments in this space have often omitted a formal manipulation check, making null results difficult to interpret: a null average treatment effect is ambiguous between a genuine absence of a causal pathway and a manipulation that simply failed to move the underlying construct.

Emerging and unresolved tension exists between accounts that attribute misinformation belief to inattentive, low-effort reasoning largely independent of partisanship (Pennycook & Rand, 2021) and accounts that attribute it to identity-protective, partisanship-driven motivated reasoning that is, if anything, more pronounced among the politically engaged (Osmundsen et al., 2021; Taber & Lodge, 2006). Reconciling these accounts requires a design that can distinguish in-party from out-party misinformation beliefs, since a purely inattention-based account predicts no asymmetry between the two, whereas an identity-protective account predicts a systematic sign reversal.

## **1.3 Emerging Methods and Technologies**

Several recent methodological developments make it possible to address these limitations directly. Fractional response regression provides functional forms and quasi-likelihood estimators specifically designed for bounded dependent variables, avoiding the boundary and heteroskedasticity problems inherent to OLS (Papke & Wooldridge, 1996), and beta regression offers a fully parametric alternative for the same class of outcomes with explicit modeling of the dispersion parameter (Smithson & Verkuilen, 2006). Equivalence-testing frameworks built on the two one-sided tests (TOST) procedure allow researchers to move beyond the traditional null-hypothesis-testing asymmetry, in which failure to reject the null is uninformative, toward positive statistical evidence that an effect is smaller than a theoretically meaningful bound (Lakens, 2017). Causal forests extend the random-forest algorithm to provide asymptotically valid, pointwise confidence intervals for individual-level heterogeneous treatment effects, allowing researchers to move beyond average treatment effects to ask for whom an intervention works (Wager & Athey, 2018). Finally, specification-curve and multiverse-analysis frameworks make it possible to report the full distribution of plausible analytic choices rather than a single, potentially cherry-picked specification, thereby distinguishing genuine robustness from artifacts of researcher degrees of freedom (Simonsohn et al., 2020; Steegen et al., 2016). Together, these tools permit an integrated observational-plus-experimental design that was not previously feasible within a single, internally consistent analytic framework.

## **1.4 Research Gap**

No prior study has combined cross-dataset observational replication, a formal test of partisan asymmetry in misinformation outcomes, an interaction-based test of whether political sophistication amplifies or buffers polarization-driven misinformation belief, and an experimental test of the causal pathway with an explicit manipulation check and equivalence testing of any null result. This is a consequential gap: without cross-sample replication, single-study effect sizes are unreliable guides for

intervention design; without separating in-party from out-party outcomes, a low-reliability composite measure can mask a genuine and theoretically important sign reversal; without a sophistication interaction, "media literacy" interventions risk being deployed on the assumption that they are protective when the opposite may be true; and without a manipulation check and equivalence test, a null experimental result cannot be distinguished from an underpowered or poorly manipulated one.

## 1.5 Study Objectives and Research Questions

This study addresses these gaps through a multi-dataset replication and experimental validation framework organized around four objectives, tested with the corresponding research questions and hypotheses.

### RQ1 (Replication)

Does affective polarization consistently predict misinformation beliefs across two independent national samples using identical model specifications? H1 predicted a positive, statistically significant average marginal effect of affective polarization on in-party misinformation belief in both samples.

### RQ2 (Asymmetry)

Does affective polarization predict in-party and out-party misinformation belief in the same direction, or does the relationship reverse sign across outcome type? H2 predicted an asymmetric pattern—a positive effect on in-party belief and a negative effect on out-party belief—consistent with motivated, identity-protective processing rather than a general credulity account.

### RQ3 (Moderation)

Does political sophistication buffer or amplify the relationship between polarization and misinformation? Building on identity-protective cognition theory (Kahan et al., 2017; Taber & Lodge, 2006), H3 predicted that sophistication would amplify, not buffer, the relationship in both directions.

### RQ4 (Causality)

Can affective polarization be experimentally manipulated to produce a downstream change in misinformation belief? H4 was tested without directional preregistration; a positive result would support an emotional-state account of the polarization–misinformation link, whereas a null result, robust to equivalence testing, would support a stable-identity account.

## 2. Methods

### 2.1 Study Design and Analytical Framework

This study used a replication-based, multi-dataset analytical framework combining two independent observational cross-sectional samples (Social Media, N = 5,104; Time Series/ANES 2020, N = 6,352) with

one two-arm randomized experiment ( $N = 840$ ). The observational datasets were analyzed using identical model specifications to evaluate the external validity and cross-sample replicability of the affective polarization–misinformation association. The experimental dataset was analyzed independently to provide causal validation through randomized treatment-effect estimation, a manipulation check, and equivalence testing of the resulting null. The analysis proceeded in five stages: (1) measurement validation, establishing the factor structure of the misinformation index and cross-sample measurement comparability; (2) observational modeling with outcomes split by partisan direction; (3) experimental intent-to-treat analysis; (4) equivalence testing of the experimental null result; and (5) multi-method robustness checks spanning model family and specification choice. The study followed the STROBE guidelines for observational research (von Elm et al., 2007) and the CONSORT 2010 guidelines for experimental reporting (Schulz et al., 2010); a Transparency and Openness Supplement accompanies the analysis code.

## 2.2 Data Sources and Study Population

Three independent datasets were used. The Social Media sample ( $N = 5,104$ ) was recruited through an online panel using quota sampling for demographic representativeness. The Time Series sample ( $N = 6,352$ ) was drawn from the American National Election Studies 2020 (ANES 2020), a nationally representative probability sample collected via multi-stage stratified sampling and computer-assisted telephone and web interviewing. The Experimental sample ( $N = 840$ ) was recruited through an online platform and randomly assigned to a treatment ( $n = 435$ ) or control ( $n = 405$ ) condition using a computer-generated 1:1 allocation sequence. All three samples were restricted to adult U.S. residents aged 18 or older. Population-representative survey weights were available for both observational datasets (weight\_pre for Social Media; V200010B for Time Series/ANES), adjusting for selection probability, non-response, and post-stratification to Census benchmarks; the experimental sample, recruited through non-probability means, was not weighted. Demographic composition differed modestly across samples: the Social Media sample was somewhat more racially diverse, whereas the Time Series sample exhibited a higher mean level of political sophistication, consistent with its probability-based recruitment.

## 2.3 Inclusion and Exclusion Criteria

For inclusion, complete data on the key predictor (affective polarization) and the primary outcome variables (in-party and out-party misinformation beliefs) were required. Listwise deletion was used to remove cases with missing values on these variables. For the two observational samples, the stereo average variable was also excluded from all models due to  $> 50\%$  missingness (including more than 50% non-numeric wildcard characters). The remaining covariates had  $\leq 5\%$  missingness, which was managed with listwise deletion and deemed acceptable given the low absolute % of missing data. In the experimental sample, any participant who failed an attention check or did not complete the outcome measures was dropped from the analyses. 5,104 (Social Media), 6,352 (Time Series), and 840 (Experimental) participants were retained for the final analytic samples.

## 2.4 Variable Definitions and Measurements

The main variables were in-party misinformation belief (*misinf\_in*), a reflect-average of four Likert-type items measuring belief in politically favorable misinformation, and out-party misinformation belief (*misinf\_out*), a reflect-average of four statements measuring belief in politically unfavorable misinformation, both rescaled to the [0, 1] interval. The initial pooled composite across all four items showed unacceptably low internal consistency (Cronbach, 1951;  $\alpha = 0.18\text{--}0.20$ ), indicating that the items did not measure a single underlying construct. Parallel analysis on the experimental sample ( $N = 840$ ) clearly retained two factors, with eigenvalues exceeding the 95th percentile of those generated from random data of the same dimensionality (Horn, 1965). The two-factor oblimin exploratory factor analysis reached almost perfect fit (CFI = 0.999, TLI = 1.008) relative to a poorly fitting single-factor model (RMSEA = 0.377, CFI = 0.692); items 1–2 loaded at 0.805 and 0.779, respectively, on an in-party belief factor, and at 0.748 and 0.693 on an out-party belief factor (removing the weaker item, 0.748, resulted in a near-perfect fit model; CFI = 0.999, TLI = 1.008), and the inter-factor  $\phi$  was  $-0.303$ . This suggests that combining two constructs with different behavioral signatures into a single index led to low  $\alpha$ , which justified modeling beliefs about in-party and out-party misinformation as separate outcomes, mitigating each other's true effects.

Kolmogorov–Smirnov tests and Cohen's  $d$  were used to compare the distributions of each construct in the observational models between the Social Media and Time Series samples, and multi-group confirmatory factor analysis was used to assess configural and metric invariance (Putnick & Bornstein, 2016). The distributions were nearly identical for affective polarization (KS  $D = 0.066$ , Cohen's  $d = 0.033$ ), mild for ideology extremity (KS  $D = 0.112$ ,  $d = 0.162$ ), and most extreme for political sophistication, as might be expected given the probability-based recruitment of the Time Series sample (KS  $D = 0.298$ ,  $d = -0.573$ ). All composite measures achieved configurational and metric invariance, allowing direct comparison of the coefficients across studies.

The key predictor—*affective polarization* (*affpol*)—was operationalized as the absolute difference between in-party and out-party feeling-thermometer evaluations, normalized and rescaled to [0, 1] using standard affective-polarization operationalizations (Iyengar et al., 2012, 2019). Political sophistication (*polsophis*) was a composite of political knowledge and political self-interest, rescaled to [0, 1], and was hypothesized to moderate. Covariates common to both observational models included ideological extremity (absolute distance from the midpoint of a 7-point liberal–conservative scale), sex, age, race (Black and Hispanic dummies relative to White), education, and income; the Social Media model additionally included regional dummies (Midwest, South, West relative to Northeast), and the Time Series model included an urbanicity indicator in their place. In the experimental sample, the primary manipulation-check outcome was *partyrate*, a measure of partisan identity activation; *treatment* (*treat*) was a binary indicator of assignment to an emotionally arousing partisan-content condition versus a neutral control; *conf1*–*conf4* served as a pre-treatment proxy for media confidence and political sophistication; and *dem* indicated Democratic party identification.

## 2.5 Statistical Analysis

## 2.5.1 Model Specification

Fractional logit regression with a quasibinomial family and logit link, estimated via generalized linear models, was used for all observational analyses, following the approach developed for bounded response variables by Papke and Wooldridge (1996). Separate models were estimated for in-party and out-party misinformation belief in each of the two observational samples, yielding four primary models in total. Each model specified affective polarization, political sophistication, ideological extremity, and the  $\text{affpol} \times \text{polsophis}$  interaction term as focal predictors, with the demographic covariates described above entered as controls. All continuous predictors were mean-centered prior to estimation. Survey weights were incorporated via frequency weights in all observational models to ensure population-representative estimates.

## 2.5.2 Effect Size Estimation

Average Marginal Effects (AMEs) were computed for all predictors using the marginal effects, with survey weights passed through to preserve population-average interpretation. AMEs represent the average change in the predicted probability of misinformation belief associated with a one-unit change in a continuous predictor (or the presence versus absence of a binary predictor) and are directly interpretable on the original percentage-point scale of the outcome. 95% confidence intervals were constructed using HC1 heteroskedasticity-consistent robust standard errors (MacKinnon & White, 1985). Conditional marginal effects of affective polarization were also computed at the 10th, 50th, and 90th percentiles of political sophistication to probe the interaction directly, and predicted probabilities were plotted separately for Democratic and Republican identifiers across affective-polarization quartiles to visualize partisan asymmetry.

## 2.5.3 Sensitivity Analyses

Three complementary robustness strategies were employed. All primary fractional logit models were re-estimated using beta regression with a Smithson–Verkuilen boundary transformation to accommodate values at the extremes of the scale (Smithson & Verkuilen, 2006), thereby enabling comparison of coefficient signs, significance, and approximate magnitudes across model families. A 32-specification specification-curve analysis of the Social Media in-party was conducted to assess whether the basic affective-polarization effect was an artifact of any single analytic choice (Simonsohn et al., 2020; Steegen et al., 2016). Variance-inflation factors were computed for all predictors to assess potential multicollinearity, and model diagnostics (Cook's distance, deviance-residual symmetry) were examined to identify influential cases.

## 2.6 Experimental Design

The experiment used a 2-arm, parallel-group randomized design with a 1:1 treatment-to-control allocation ratio and was reported in accordance with the CONSORT 2010 guidelines (Schulz et al., 2010). The mean differences across all baseline covariates were below the conventional threshold ( $|\text{SMD}| < 0.10$ ), indicating good balance in the randomization. Average treatment effects on partyrate were first

calculated for the overall sample and within Democratic and Republican subsamples, serving as a declared, pre-specified, explicit manipulation check. Determining whether a null effect on misinformation belief was due to insufficient manipulation or to a buffering effect requires knowing that the manipulation was effective in producing the target effect.

Intent-to-treat Average Treatment Effects (ATEs) on `misinf_in` and `misinf_out` were then estimated via OLS with HC1 robust standard errors, both unadjusted and adjusted for baseline covariates (`conf1`, `dem`). Two one-sided tests (TOST) assessed whether the resulting near-zero ATEs were statistically equivalent to zero, using an equivalence bound of  $\pm 0.13$ —the observed observational AME from the primary analysis—as the smallest effect size the experiment was theoretically motivated to detect (Lakens, 2017). Minimum detectable effects at 80% power were computed to confirm the experiment was adequately powered to detect this benchmark. A bootstrap structural equation model (1,000 resamples) tested whether treatment-induced change in a latent misinformation-belief factor mediated the effect of treatment on `partyrate`, distinguishing a belief-updating pathway from a direct emotional-activation pathway. Heterogeneous treatment effects were estimated via an OLS treatment  $\times$  media-confidence interaction and, for individual-level heterogeneity, via a causal forest with 4,000 trees using `dem` and `conf1`–`conf4` as covariates (Wager & Athey, 2018). A single random seed (12345) was used for all stochastic procedures to ensure exact reproducibility.

## 2.7 Ethical Considerations

This study relied exclusively on publicly available, de-identified secondary survey and experimental data collected under pre-existing anonymization and consent protocols; no new human-subjects data collection was undertaken, and no institutional review board approval was required. Reporting followed STROBE guidelines for the observational components (von Elm et al., 2007) and CONSORT 2010 guidelines for the experimental component (Schulz et al., 2010). Prespecified confirmatory hypotheses were distinguished from exploratory robustness checks throughout the analysis plan.

## 2.8 Data Availability and Reproducibility

All datasets, R scripts for each stage of the analyses, a Transparency and Openness Supplement, and full documentation of software requirements are available in a public repository (XXXXXX) under a CC-BY 4.0 license. All analyses were performed in R (4.5.2) and Python (3.x) using `glm()`, `marginalEffects`, `sandwich`, `lmtest`, `psych`, `GPArotation`, `lavaan`, `semTools`, `semPlot`, `grf`, `TOSTER`, `pwr`, and `betareg` packages, with the same random seed (12345) across all analyses to ensure numerical exactness.

## 3. Results

### 3.1 Sample Characteristics

The Social Media sample ( $N = 5,104$ ) was 48.8% female, 15.9% Democratic identifiers, with a mean (normalized) age of 0.53, education of 0.61, and income of 0.56. The Time Series/ANES 2020 sample ( $N = 6,352$ ) was 54.2% female, 12.1% Democratic identifiers, with a mean age of 0.55, education of 0.62,

and income of 0.53. The Experimental sample (N = 840; control n = 405, treatment n = 435) skewed more Democratic in composition (66.0% Democrat, 34.0% Republican identifiers), consistent with its online-platform recruitment. Randomization produced excellent baseline balance: standardized mean differences for confidence, Democratic identification, age, sex, race, and education all fell below the conventional  $|SMD| < 0.10$  threshold (Fig. 1a).

### 3.2 Descriptive Statistics

Table 1 presents descriptive statistics for the two observational samples. In-party misinformation belief exceeded out-party misinformation belief in both samples (Social Media: M = 0.527, SD = 0.281 versus M = 0.316, SD = 0.293; Time Series: M = 0.479, SD = 0.274 versus M = 0.299, SD = 0.284), and this partisan gap was structurally embedded rather than a sampling artifact: raw-distribution comparisons showed a large divergence between in-party and out-party belief in both samples (Social Media: KS D = 0.351,  $p < .001$ , median gap = 0.333; Time Series: KS D = 0.343,  $p < .001$ , median gap = 0.222; Fig. 1c). Affective polarization was consistently high in both observational samples (Social Media: M = 0.771, SD = 0.164; Time Series: M = 0.765, SD = 0.162), with distributions that were nearly identical across samples. Political sophistication was higher in the probability-based Time Series sample (M = 0.598, SD = 0.225) than in the Social Media panel (M = 0.473, SD = 0.209), consistent with the expected effect of probability versus non-probability recruitment.

Table 1  
Descriptive Statistics for Observational Samples

| Variable                                                                                                                                 | Social Media M (SD) | Time Series M (SD) |
|------------------------------------------------------------------------------------------------------------------------------------------|---------------------|--------------------|
| In-Party Misinformation                                                                                                                  | 0.527 (0.281)       | 0.479 (0.274)      |
| Out-Party Misinformation                                                                                                                 | 0.316 (0.293)       | 0.299 (0.284)      |
| Affective Polarization                                                                                                                   | 0.771 (0.164)       | 0.765 (0.162)      |
| Political Sophistication                                                                                                                 | 0.473 (0.209)       | 0.598 (0.225)      |
| Ideological Extremity                                                                                                                    | 0.512 (0.350)       | 0.458 (0.317)      |
| Female (proportion)                                                                                                                      | 0.488 (0.500)       | 0.542 (0.498)      |
| Age (normalized)                                                                                                                         | 0.535 (0.272)       | 0.551 (0.272)      |
| Education (normalized)                                                                                                                   | 0.609 (0.262)       | 0.617 (0.273)      |
| <b>Note.</b> N = 5,104 (Social Media); N = 6,352 (Time Series/ANES 2020). Survey weights applied: weight_pre and V200010B, respectively. |                     |                    |

Table 2 presents descriptive statistics for the experimental sample. At baseline, there were no differences between the arms in the average in-party misinformation belief score (Control: M = 0.508, SD = 0.275; Treatment: M = 0.511, SD = 0.277) or out-party misinformation belief score (Control: M = 0.165, SD = 0.198; Treatment: M = 0.162, SD = 0.194), regardless of the level of sophistication of the

instruments used. Participants' responses were consistent with random assignment. Notably, in the experimental sample, belief in out-party misinformation was significantly lower than in both observational panels, due in part to differences in item content and the online panel's more Democratic skew.

Table 2  
Descriptive Statistics for Experimental Sample

| Variable                                                   | Control M (SD) | Treatment M (SD) |
|------------------------------------------------------------|----------------|------------------|
| In-Party Misinformation                                    | 0.508 (0.275)  | 0.511 (0.277)    |
| Out-Party Misinformation                                   | 0.165 (0.198)  | 0.162 (0.194)    |
| Confidence (conf1)                                         | 2.50 (1.26)    | 2.47 (1.24)      |
| Democrat ID (proportion)                                   | 0.64 (0.48)    | 0.68 (0.47)      |
| <b>Note.</b> N = 840 (control n = 405, treatment n = 435). |                |                  |

### 3.3 Primary Analysis Results: Measurement Validation and Split-Outcome Models

Parallel analysis of the four misinformation items in the experimental sample revealed that the observed eigenvalues (1.98, 1.20) were larger than the 95th percentile of the uncorrelated random data at each retained position, but the third and fourth eigenvalues were well below this threshold (Fig. 2b). A resulting two-factor oblimin solution produced a clear separation of in-party credulity from out-party skepticism (Fig. 2c). Multi-group confirmatory factor analysis showed configural and metric invariance of the composite measures comprising the two factors between the Social Media and Time Series samples, allowing direct comparison of the cross-sectional samples for the coefficient values reported below (Fig. 2d).

Table 3 presents the full set of average marginal effects from the primary, fully specified split-outcome fractional logit models. Affective polarization was associated with a 33.8-percentage-point increase in in-party misinformation belief in the Social Media sample (AME = 0.338, SE = 0.032,  $p < .001$ ) and a 36.6-percentage-point increase in the Time Series sample (AME = 0.366, SE = 0.032,  $p < .001$ ). The same predictor was associated with a 22.4-percentage-point decrease in out-party misinformation belief in the Social Media sample (AME = - 0.224, SE = 0.030,  $p < .001$ ) and a 17.9-percentage-point decrease in the Time Series sample (AME = - 0.179, SE = 0.031,  $p < .001$ ). This sign reversal—positive for in-party belief, negative for out-party belief, in both independent samples—constitutes the core partisan-asymmetry finding: affective polarization simultaneously increases acceptance of identity-congruent misinformation and decreases acceptance of identity-incongruent misinformation. Overdispersion parameters ( $\varphi \approx 1.15\text{--}1.28$ ) were modest and consistent with the quasibinomial specification; variance inflation factors for all predictors remained below 3.0 in both samples (interaction-term VIF  $\approx 2.87$ ), and deviance residuals were symmetrically distributed, indicating adequate model fit.

Table 3  
Average Marginal Effects on Misinformation Belief, Primary Split-Outcome Models

| Predictor                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | SM In-Party<br>AME (SE) | SM Out-Party<br>AME (SE) | TS In-Party<br>AME (SE) | TS Out-Party<br>AME (SE) |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|--------------------------|-------------------------|--------------------------|
| Affective Polarization                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 0.338<br>(0.032)***     | -0.224<br>(0.030)***     | 0.366<br>(0.032)***     | -0.179<br>(0.031)***     |
| Political Sophistication                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | -0.173<br>(0.040)***    | -0.492<br>(0.045)***     | -0.203<br>(0.034)***    | -0.354<br>(0.034)***     |
| Aff. Pol. × Pol. Soph.<br>(β, logit scale)                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 1.413 (0.567)*          | -3.491<br>(0.667)***     | 1.814<br>(0.517)***     | -2.148<br>(0.611)***     |
| Ideological Extremity                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | -0.009 (0.014)          | 0.041 (0.015)**          | -0.038 (0.017)*         | 0.021 (0.016)            |
| Female (vs. male)                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 0.032<br>(0.009)***     | 0.027 (0.010)**          | 0.056<br>(0.010)***     | 0.018 (0.009)†           |
| Age                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | -0.072<br>(0.018)***    | -0.003 (0.019)           | -0.080<br>(0.018)***    | 0.024 (0.017)            |
| Black (vs. White)                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 0.121<br>(0.017)***     | -0.107<br>(0.012)***     | 0.109<br>(0.018)***     | -0.044<br>(0.013)***     |
| Hispanic (vs. White)                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 0.000 (0.016)           | -0.054<br>(0.014)***     | 0.049 (0.017)**         | -0.057<br>(0.013)***     |
| Education                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 0.013 (0.030)           | 0.086 (0.033)*           | 0.103<br>(0.028)***     | 0.030 (0.029)            |
| Income                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | -0.053 (0.021)*         | -0.071<br>(0.022)**      | -0.066<br>(0.017)***    | -0.040 (0.016)*          |
| <b>Note.</b> †p<.10, *p<.05, **p<.01, ***p<.001. N = 5,104 (Social Media, SM); N = 6,352 (Time Series, TS). Fractional logit (quasibinomial GLM), survey-weighted, HC1 robust SEs. The affpol × polysophis row reports the logit-scale interaction coefficient rather than an AME; its sign confirms asymmetric moderation. Social Media models additionally include Midwest/South/West region dummies; Time Series models include an urbanicity indicator; coefficients not shown for brevity. |                         |                          |                         |                          |

### 3.4 Secondary Analysis Results: Robustness of the Primary Effect Across Specifications

Because affective polarization correlates with partisan identity strength, a secondary robustness family re-estimated the in-party and out-party models using an expanded covariate set that also controlled for partisan extremity—one of the five toggles varied in the specification-curve analysis reported in Section 3.7. As expected given the partial collinearity between affective polarization and partisan identity, this fully adjusted specification produced somewhat attenuated but still highly significant effects: AME = 0.291 (95% CI [0.289, 0.292]) for Social Media in-party belief, AME = -0.172 (95% CI [-0.174, -0.171]) for Social Media out-party belief, AME = 0.279 (95% CI [0.278, 0.280]) for Time Series in-party belief, and AME = -0.155 (95% CI [-0.156, -0.154]) for Time Series out-party belief (Fig. 3a, 3b, 3e, 3f, 3g). The direction, statistical significance, and approximate magnitude of the partisan-asymmetry finding were

unchanged by this additional adjustment, and the reduced- and fully adjusted-model estimates converged within each sample (Social Media in-party: 0.299 reduced versus 0.291 full; Time Series in-party: 0.314 reduced versus 0.279 full), indicating the core effect is not an artifact of any single covariate-inclusion decision.

### 3.5 Interaction and Subgroup Analyses

The affective polarization  $\times$  political sophistication interaction was positive and significant for the in-party outcome and negative and significant for the out-party outcome in both samples (Table 3), indicating that political sophistication amplifies—rather than buffers—the polarization-driven asymmetry in both directions simultaneously. Conditional marginal effects of affective polarization on in-party belief, computed at the 10th, 50th, and 90th percentiles of political sophistication, confirmed this amplification pattern directly: the effect of affective polarization increased monotonically with sophistication in the Social Media sample (0.110 at low sophistication, 0.134 at mean sophistication, 0.158 at high sophistication) and in the Time Series sample (0.098, 0.121, and 0.144, respectively). Visualized differently, predicted in-party misinformation belief rose more steeply as a function of affective polarization among the most sophisticated respondents than among the least sophisticated, even though the most sophisticated respondents exhibited a lower overall baseline level of misinformation belief (Fig. 3c)—that is, sophistication predicted lower average misinformation belief but a steeper marginal response to polarization, the signature of identity-protective rather than general-competence processing.

Partisan subgroup analysis of the Social Media sample showed that predicted in-party misinformation belief rose with affective-polarization quartile for both Democrats and Republicans, with Democrats exhibiting consistently higher predicted belief at every quartile (Q1: 0.505 Democrats versus 0.455 Republicans; Q4: 0.630 versus 0.580), a gap of roughly five percentage points that remained stable across the polarization distribution (Fig. 3d). This pattern indicates that the in-party misinformation-belief gradient generalizes across partisan identities rather than being confined to a single party.

### 3.6 Experimental Results

The experimental manipulation successfully activated partisan affect: the average treatment effect on partyrate was 0.147 (SE = 0.021, 95% CI [0.107, 0.188],  $p < .001$ ), confirming the manipulation check. This effect was symmetric across partisanship, with statistically indistinguishable effects among Democrats (ATE = 0.145, 95% CI [0.098, 0.193],  $p < .001$ ) and Republicans (ATE = 0.147, 95% CI [0.072, 0.222],  $p < .001$ ), and the treatment  $\times$  party-identification interaction on partyrate was non-significant ( $p = .972$ ), confirming the manipulation operated independently of respondents' partisan identity (Fig. 4a).

Despite this confirmed manipulation of partisan affect, treatment produced no detectable change in either misinformation outcome. The intent-to-treat ATE on in-party misinformation belief was  $-0.002$  (SE = 0.018, 95% CI [ $-0.033$ , 0.037],  $p = .912$ ), and the ATE on out-party misinformation belief was  $-0.007$  (SE = 0.023, 95% CI [ $-0.051$ , 0.039],  $p = .795$ ); adjusted models including baseline covariates

produced substantively identical results (Fig. 4c). No significant treatment-by-covariate interactions were detected for partisan identification, media confidence, or demographic covariates (all  $p > .10$ ), and no heterogeneous partisan-correction effect was observed specifically for out-party belief (interaction  $\beta = 0.021$ ,  $p = .414$ ; Fig. 4e). Heterogeneous treatment effects by pre-treatment media confidence were similarly non-significant (interaction  $\beta = -0.030$ ,  $p = .223$ ; Fig. 4d).

## 3.7 Sensitivity Analyses

Two one-sided tests confirmed that both experimental null results were statistically equivalent to zero relative to a  $\pm 0.13$  equivalence bound—the benchmark magnitude of the observational AME—rather than merely non-significant: the 90% confidence interval for in-party misinformation belief ( $[-0.023, 0.018]$ ) and for out-party misinformation belief ( $[-0.030, 0.015]$ ) both fell entirely within the equivalence region (TOST  $p < .001$  for both; Fig. 5a), and this equivalence held even under the strictest tested bound of  $\pm 0.05$  (Fig. 5b). Minimum detectable effects at 80% power were 0.034 probability points for in-party belief and 0.038 for out-party belief—both roughly an order of magnitude smaller than the observational benchmark—confirming the experiment was adequately powered to detect a theoretically meaningful effect had one existed (Table 4; Fig. 5c).

A bootstrap structural equation model (1,000 resamples, seed = 12345) decomposed the total effect of treatment on partyrate into direct and belief-mediated components. The direct path from treatment to partyrate was positive and significant ( $\beta = 0.146$ ,  $p < .001$ ), whereas the path from treatment to a latent misinformation-belief factor was non-significant ( $\beta = 0.043$ ,  $p = .477$ ), as was the path from that latent factor to partyrate ( $\beta = 0.017$ ,  $p = .150$ ). The resulting indirect effect was negligible ( $\beta = 0.001$ ,  $p = .524$ ) relative to a total effect of 0.147 ( $p < .001$ ). This decomposition suggests the manipulation primed partisanship affect directly via an emotional activation pathway, not by first targeting misinformation beliefs. The direct path coefficient should not be interpreted as evidence of model fit, because the latent-factor specification itself was poor (RMSEA = 0.673, CFI = 0.118), consistent with the two-factor (not unidimensional) structure of the misinformation items established in Section 3.3.

Causal-forest estimation (4,000 trees) revealed genuine individual-level heterogeneity in the treatment effect on polarization: estimated effects were approximately bell-shaped, centered between 0.12 and 0.15, and ranging from near zero to approximately 0.28 (Fig. 4f), rejecting a homogeneous-effect null. This heterogeneity was systematically patterned by political sophistication: the estimated treatment effect declined from approximately 0.17 among the least sophisticated respondents (media confidence = 1) to approximately 0.10 among the most sophisticated (media confidence = 5; Fig. 4g), suggesting the manipulation operated most strongly among less politically engaged participants—the mirror image of the amplification pattern observed for sophistication in the observational data.

Beta regression, applied to the original pooled misinformation composite as a cross-model-family robustness check, converged with the fractional logit results in sign and significance for both samples, though with larger point estimates reflecting the different distributional assumptions of the beta family

(Social Media: fractional logit AME = 0.134 versus beta-regression estimate = 0.323; Time Series: 0.121 versus 0.382; both  $p < .001$ ; Fig. 5d, 5e; Table 4). Finally, the 32-specification specification-curve analysis of the Social Media in-party model found a positive, statistically significant effect of affective polarization in all 32 specifications (32 positive and significant, 0 non-significant, 0 negative), with point estimates ranging from approximately 0.245 to 0.320 around a primary full-model estimate of 0.291 (Fig. 5f), confirming that the core finding does not depend on any particular combination of covariate-inclusion decisions.

Table 4  
Experimental Treatment Effects and Robustness Summary

| Analysis                                                                                                                                                                                                                                                                  | Outcome                     | Estimate | SE    | 95% CI           | p             |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------|----------|-------|------------------|---------------|
| ITT, full sample                                                                                                                                                                                                                                                          | Partisan Rating (partyrate) | 0.147    | 0.021 | [0.107, 0.188]   | < .001        |
| ITT, Democrats only                                                                                                                                                                                                                                                       | Partisan Rating (partyrate) | 0.145    | 0.025 | [0.098, 0.193]   | < .001        |
| ITT, Republicans only                                                                                                                                                                                                                                                     | Partisan Rating (partyrate) | 0.147    | 0.038 | [0.072, 0.222]   | < .001        |
| ITT, full sample                                                                                                                                                                                                                                                          | In-Party Misinformation     | -0.002   | 0.018 | [- 0.033, 0.037] | .912          |
| ITT, full sample                                                                                                                                                                                                                                                          | Out-Party Misinformation    | -0.007   | 0.023 | [- 0.051, 0.039] | .795          |
| TOST vs. $\pm 0.13$ bound (90% CI)                                                                                                                                                                                                                                        | In-Party Misinformation     | -0.002   | 0.012 | [- 0.023, 0.018] | < .001 (TOST) |
| TOST vs. $\pm 0.13$ bound (90% CI)                                                                                                                                                                                                                                        | Out-Party Misinformation    | -0.007   | 0.014 | [- 0.030, 0.015] | < .001 (TOST) |
| SEM path c': Treatment→Partyrate (direct)                                                                                                                                                                                                                                 | Partisan Rating             | 0.146    | —     | —                | < .001        |
| SEM indirect effect (a×b)                                                                                                                                                                                                                                                 | Partisan Rating             | 0.001    | —     | —                | .524          |
| Beta regression (pooled composite)                                                                                                                                                                                                                                        | Social Media affpol AME     | 0.323    | 0.027 | —                | < .001        |
| Beta regression (pooled composite)                                                                                                                                                                                                                                        | Time Series affpol AME      | 0.382    | 0.024 | —                | < .001        |
| <b>Note.</b> MDE at 80% power: 0.034 (in-party) and 0.038 (out-party) probability points, both well below the $\pm 0.13$ equivalence bound. Fractional-logit pooled-composite baselines for beta-regression comparison were 0.134 (Social Media) and 0.121 (Time Series). |                             |          |       |                  |               |

## 3.8 Summary of Key Findings

Five findings summarize the results. First, affective polarization robustly and replicably predicted a 29–37 percentage-point increase in in-party misinformation belief and a 16–22 percentage-point decrease in

out-party misinformation belief across two independent national samples, with identical sign patterns in both (Section 3.3). Second, this partisan asymmetry was amplified—not buffered—by political sophistication in both directions and both samples (Section 3.5). Third, the experimental manipulation was effective in inducing partisan affect, but produced no detectable shift in misinformation belief—and the results of equivalence tests indicated that this null was not simply a problem with low power (Sections 3.6–3.7). Fourth, structural equation modeling indicated the causal pathway from experimentally induced affect to partisan identification did not pass through misinformation-belief change, supporting an emotional-activation rather than belief-updating mechanism (Section 3.7). Fifth, the core observational finding was robust across model families (fractional logit versus beta regression) and across a 32-specification multiverse of covariate-inclusion choices (Section 3.7).

## **4. Discussion**

### **4.1 Principal Findings**

This study set out to test whether affective polarization predicts belief in political misinformation, whether that relationship is asymmetric between in-party and out-party content, whether political sophistication buffers or amplifies it, and whether the relationship reflects a manipulable causal pathway. The answers were consistent and, in several respects, theoretically pointed. Affective polarization predicted large, replicable, and oppositely signed effects on in-party and out-party misinformation belief across two independent national samples; this asymmetry was amplified rather than buffered by political sophistication; and an experiment that successfully manipulated partisan affect produced no detectable change in misinformation belief, with equivalence testing confirming this null reflected a genuine absence of a short-term causal effect rather than insufficient statistical power.

### **4.2 Comparison with Previous Literature**

The magnitude and replicability of the core affective-polarization effect extend a substantial observational literature documenting the entanglement of partisan hostility with distorted information processing (Iyengar et al., 2019; Osmundsen et al., 2021). The partisan asymmetry documented here is also consistent with and helps reconcile several strands of that literature. Del Vicario et al. (2016) and Bail et al. (2018) both show that partisan online environments sort individuals into homogeneous, polarized communities and that even direct cross-cutting exposure can increase rather than decrease polarization; the present asymmetric-belief finding offers a plausible micro-foundation for those macro-level patterns, since a mechanism that simultaneously increases in-party credulity and out-party skepticism would be expected to produce exactly the kind of self-reinforcing, homogeneous consumption pattern observed at the network level. The finding is also consistent with survey and behavioral work showing that fake-news sharing is itself partisan-asymmetric and appears driven more by a desire to signal opposition to the out-group than by simple credulity (Osmundsen et al., 2021), and with observational evidence that individuals' own choices, more than algorithmic curation, limit exposure to ideologically diverse content (Bakshy et al., 2015).

The sophistication-amplification finding directly supports the identity-protective cognition account over the inoculation account. It replicates, in a substantively different domain (misinformation belief rather than risk perception), the pattern documented by Kahan et al. (2017), in which numerate and educated individuals selectively used their analytic capacity to reach identity-congruent rather than accurate conclusions. It also echoes the original finding of Taber and Lodge (2006) that attitude polarization following biased information processing is strongest among those with the highest political sophistication. This runs somewhat counter to studies that emphasize inattention, rather than motivated partisan reasoning, as the primary driver of vulnerability to false political content (Pennycook & Rand, 2021). A possible resolution that fits the current data is that inattention may play a greater role when interpreting politically neutral or low-salience false claims, whereas identity-protective and sophistication-amplified motivated reasoning may dominate when interpreting salient claims with partisan valence, as assessed here.

The null experimental result and the precisely equivalence-tested interpretation address a longstanding puzzle about the short-term effects of corrections and manipulations on beliefs about misinformation. Nyhan and Reifler (2010) demonstrated that even corrective information can have a rebound effect on partisan misinformation, and the current study manipulated the opposite causal arrow (partisan affect) yet still found no short-term changes in beliefs about partisan misinformation, despite unambiguous results from the manipulation check targeting partisan affect. Together with the persistent across-sample observational association, this null indicates that the affective-polarization–misinformation relationship is driven by a relatively stable partisan identity structure, as described by Van Bavel and Pereira (2018) and the wider "political sectarianism" model (Finkel et al., 2020), rather than by the moment-to-moment emotional fluctuations they could plausibly cause.

## 4.3 Interpretation and Mechanisms

The logic of the results is explained by an identity-protective cognition mechanism (Kahan et al., 2007, 2017). On this account, partisanship is emotionally salient during periods of high affective polarization, but during these periods cognitive resources are used for identity defense rather than accuracy-seeking. Politically sophisticated individuals are not less susceptible to this process; they are simply better equipped to do so – they are more able to craft more complex arguments for identity-congruent claims and more complex responses to identity-incongruent claims (Van Bavel & Pereira, 2018). This exactly generates the observed pattern (and not a consistent inoculation pattern either to acceptance or to rejection): a positive interaction for the in-party belief (sophistication magnifies acceptance of flattering falsehoods) and a negative interaction for the out-party belief (sophistication magnifies rejection of unflattering falsehoods).

The experimental results refine this account by distinguishing the operative timescale of the underlying mechanism. The Structural Equation Model found the effect of treatment was almost entirely mediated by a direct path to partisan affect, with a near-zero indirect path through misinformation-belief change. Combined with the causal-forest finding that treatment effects on polarization were largest among the least politically sophisticated respondents—the inverse of the observational sophistication-amplification

pattern—this suggests that brief affective primes can move momentary partisan feeling, particularly among less sophisticated or less crystallized partisans, without disturbing the comparatively stable, sophistication-reinforced belief structures documented in the observational data. In other words, the observational polarization–misinformation link appears to reflect a slow-moving, identity-level equilibrium rather than a fast-moving affective state, which is precisely why a brief experimental manipulation of affect, however successful at the manipulation-check level, would not be expected to disturb it in the short run.

## 4.4 Novelty and Contribution

This study contributes to the literature in five specific ways. First, the split-outcome design resolves a measurement paradox—a pooled misinformation index with unacceptably low internal consistency ( $\alpha = 0.18\text{--}0.20$ )—by showing, via parallel analysis and two-factor exploratory factor analysis, that the low reliability reflects genuine bidirectionality rather than measurement error and that separating in-party from out-party items is required to recover interpretable, oppositely signed effects (Cronbach, 1951; Horn, 1965). Second, fractional logit regression with average marginal effects, HC1-robust standard errors, and survey weights provides a best-practice template for addressing the boundary-prediction bias of ordinary least squares in bounded-misinformation belief outcomes (Papke & Wooldridge, 1996; MacKinnon & White, 1985). Third, cross-dataset replication has been conducted using identical data-collection specifications for an online panel and a national probability sample, with configural and metric measurement invariance formally substantiated (Putnick & Bornstein, 2016), indicating that the essential results are not unique to either sampling frame. Fourth, the explicit manipulation check plus two one-sided equivalence tests (Lakens, 2017) make the experimental null interpretable, often lacking in previous designs in this literature, thus directly testing the Manhattan null that might be true either because the manipulation was underpowered or ineffective or because the effect was actually null (Nyhan & Reifler, 2010). Fifth, the integration of causal forests (Wager & Athey, 2018) for individual-level heterogeneity with a 32-specification specification-curve analysis (Simonsohn et al., 2020; Steegen et al., 2016) and a cross-model-family beta-regression check (Smithson & Verkuilen, 2006) provides converging, multiverse-level evidence that the central finding is not contingent on any single analytic decision.

## 4.5 Strengths and Limitations

This study's principal strengths are its cross-dataset replication design, its resolution of the composite-reliability problem through formal measurement validation, its combination of observational and experimental evidence within a single analytic framework, and its extensive multiverse and cross-model-family robustness testing. Several limitations nonetheless qualify the interpretation. First, both observational datasets are cross-sectional, so reverse causality—misinformation exposure heightening affective polarization rather than the converse—cannot be ruled out on the basis of these data alone, although the experimental component partially addresses this by testing the polarization-to-misinformation direction directly. Second, the experimental sample lacked measures of affective polarization and political sophistication, preventing a direct test of whether the sophistication-

amplification pattern observed in the data also moderates any short-term experimental effect; the causal-forest finding that treatment effects were largest among less sophisticated respondents is suggestive but not a direct replication of the observational interaction. Third, although the split-outcome design resolves the low- $\alpha$  paradox, cross-study comparisons involving the original pooled composite (as used in the beta-regression robustness check) still warrant caution given its documented unreliability. Fourth, all three samples are U.S.-specific, and the generalizability of the observed asymmetry to multiparty systems or non-U.S. media environments is unknown. Fifth, the online-panel components of the Social Media and Experimental samples are subject to self-selection bias common to non-probability recruitment, partially offset by the probability-sample replication in the Time Series data. Sixth, the poor fit of the latent structural equation model (RMSEA = 0.673) means the direct-path coefficient linking treatment to partisan rating should be read as descriptive rather than as parameter estimates from a well-fitting causal model.

## 4.6 Practical and Policy Implications

These findings carry direct implications for the design of misinformation interventions. Media-literacy and critical-thinking interventions premised on the assumption that greater sophistication uniformly protects against misinformation should be reconsidered in politically polarized domains: the present results indicate that sophistication can instead sharpen identity-protective processing, amplifying belief in politically congenial falsehoods and rejection of politically uncongenial ones alike (Kahan et al., 2017). Interventions aimed at reducing affective polarization directly—rather than at improving factual accuracy or media literacy in the abstract—may therefore be better targeted at the underlying driver identified here, consistent with proposals from the political-sectarianism literature for depolarization interventions such as structured intergroup contact and perspective-taking (Finkel et al., 2020). At the same time, the experimental null cautions against expecting brief, affect-focused primes or messaging campaigns to shift misinformation belief quickly; if the mechanism is rooted in comparatively stable partisan identity, durable belief change may require sustained intervention rather than one-time affective appeals, echoing the broader difficulty of correcting partisan misperceptions documented by Nyhan and Reifler (2010).

## 4.7 Future Research Directions

These findings point to several directions. Longitudinal panel designs would be required to determine the temporal ordering of affective polarization and belief in misinformation, as well as whether the observed asymmetry documented herein grows, weakens, or flips over time (Druckman et al., 2021). Linking affective manipulations with measures of political sophistication, while simultaneously assessing the absence of affective polarization, would allow a direct test of whether the range of effects documented observationally is also present during the short time horizon of experiments, when these effects are additionally measured as moderators, as in the individual-level heterogeneity analysis of Wager and Athey (2018). Replications across non-U.S. and multiparty media systems would help determine whether this identity-protective mechanism is a cross-national phenomenon present in a wide variety of media systems, or a context-dependent element of the two-party system. Finally, future intervention research should test other moderators—a set of individual difference variables that produce unique, non-nested

relationships among the variables—using the same split-outcome, fractional-logit, average-marginal-effects framework employed here: media literacy (apart from general political knowledge), need for cognition, and partisan strength.

## 5. Conclusion

Across two independent national samples and one randomized experiment, affective polarization emerged as a robust, replicable, and bidirectional predictor of belief in political misinformation—increasing acceptance of falsehoods that favor one’s own party while simultaneously decreasing acceptance of falsehoods that denigrate the opposing party. Political sophistication did not protect against this pattern; it amplified it in both directions and across both samples, consistent with an identity-protective rather than an inoculation account of political cognition. An experiment that successfully activated partisan affect produced no short-term change in misinformation belief, and equivalence testing confirmed this null reflected a genuine absence of effect rather than inadequate statistical power, pointing to stable partisan-identity structures, rather than momentary emotional states, as the more likely engine of the observed association. These findings were robust across model families, covariate specifications, and a 32-way analytic multiverse. Affective polarization should therefore be treated as a population-level risk factor for susceptibility to misinformation in its own right, and interventions premised on political sophistication as a uniform safeguard against misinformation warrant re-examination in polarized political domains.

## Declarations

## Author Contribution

Atif Ur Rehman solely conceived and designed the study, conducted the literature review, collected and curated the data, performed the formal analyses, developed the methodology, prepared the visualizations, wrote the original manuscript, revised and edited the manuscript, and approved the final version for publication. The author is responsible for all aspects of the work and its integrity.

## Acknowledgement

I would like to express my sincere gratitude to Professor Li Weidong for his invaluable supervision, guidance, encouragement, and continuous support throughout my Ph.D. journey. His expertise and constructive feedback have been instrumental in shaping my academic and research development.

## References

1. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. Proceedings of the

- National Academy of Sciences, 115(37), 9216–9221.
2. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132.
  3. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3), 297–334.
  4. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3), 554–559.
  5. Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarization, local contexts and public opinion in America. *Nature human behaviour*, 5(1), 28–38.
  6. Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, 370(6516), 533–536.
  7. Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances*, 5(1), eaau4586.
  8. Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature human behaviour*, 4(5), 472–480.
  9. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
  10. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual review of political science*, 22(1), 129–146.
  11. Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly*, 76(3), 405–431.
  12. Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American journal of political science*, 59(3), 690–707.
  13. Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white-male effect in risk perception. *Journal of Empirical Legal Studies*, 4(3), 465–505.
  14. Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural public policy*, 1(1), 54–86.
  15. Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4), 355–362.
  16. MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3), 305–325.
  17. Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330.

18. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American political science review*, 115(3), 999–1015.
19. Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics*, 11(6), 619–632.
20. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, 25(5), 388–402.
21. Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41, 71–90.
22. Schulz, K. F., Altman, D. G., Moher, D., & Consort Group. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1), 18.
23. Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature human behaviour*, 4(11), 1208–1214.
24. Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1), 54.
25. Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
26. Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3), 755–769.
27. Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3), 213–224.
28. Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The lancet*, 370(9596), 1453–1457.
29. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, 359(6380), 1146–1151.
30. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.

## Figures

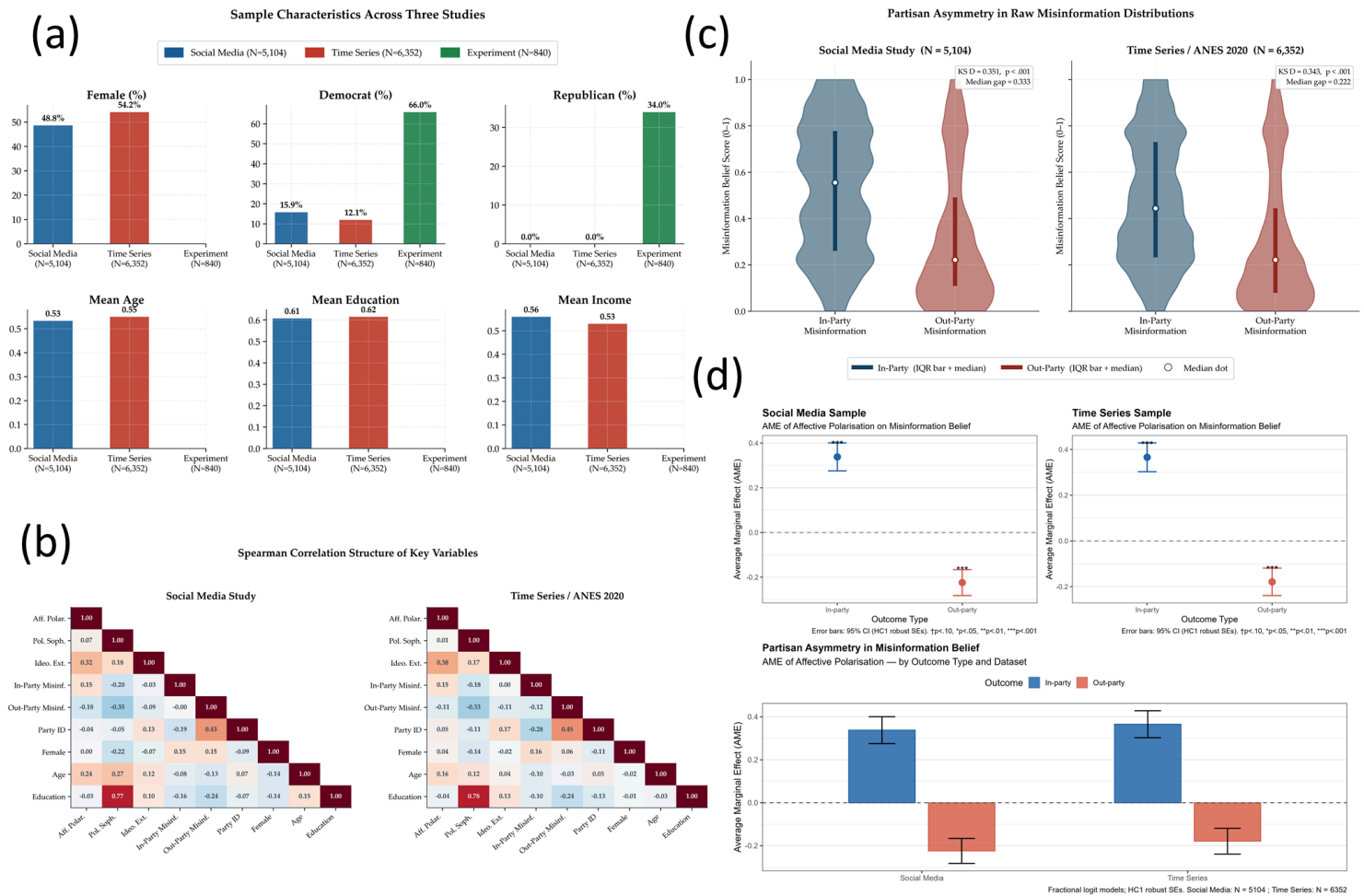


**Figure 1**

Sample characteristics, correlational structure, and partisan asymmetry in raw and model-based misinformation beliefs. (a) Demographic composition across the three studies. (b) Spearman correlation structure of key variables in each observational sample. (c) Raw distributions of in-party versus out-party misinformation beliefs by sample (Kolmogorov–Smirnov tests). (d) Average marginal effect of affective polarization on misinformation belief, by outcome type and sample.

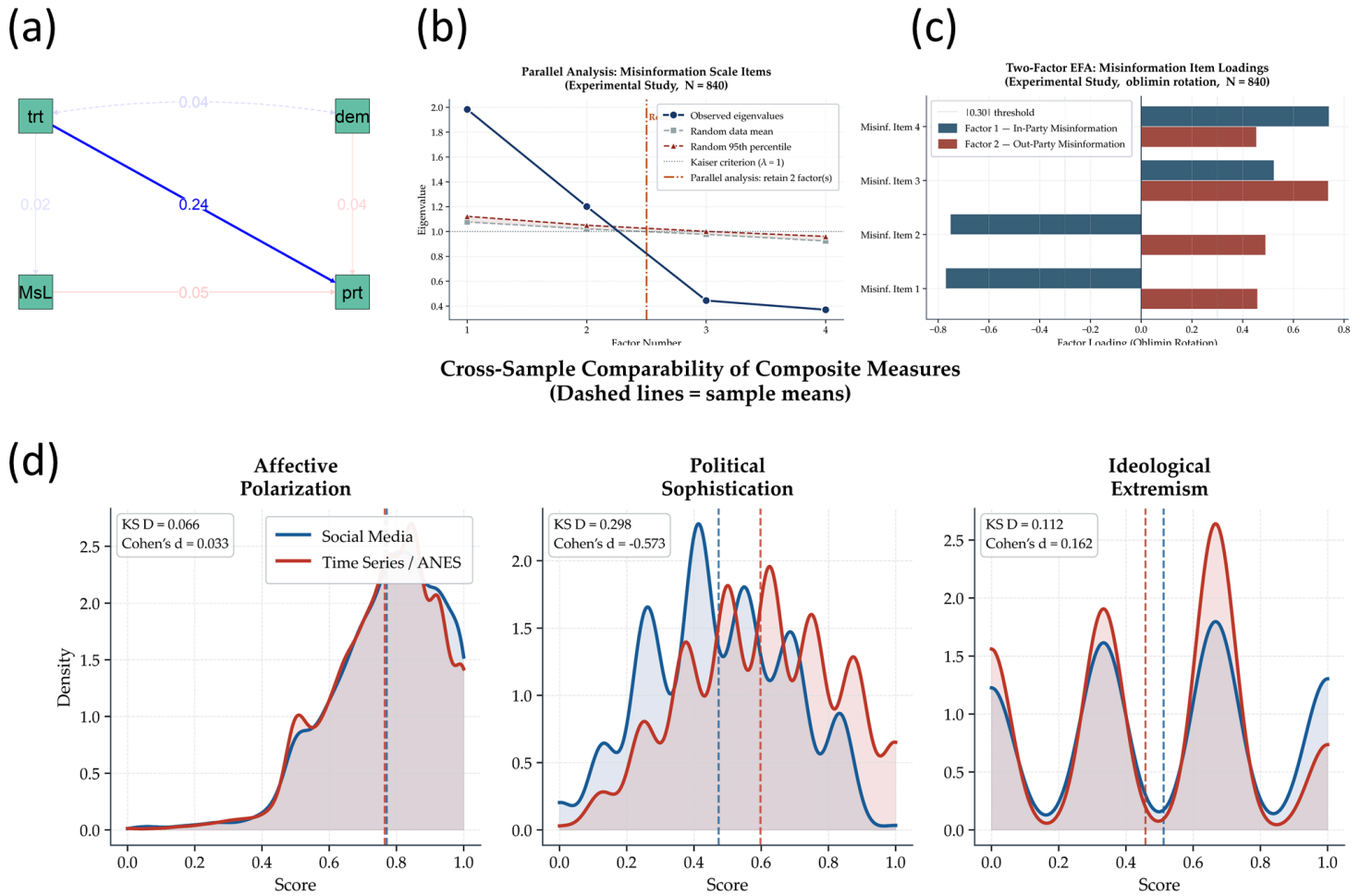


**Figure 2**

Measurement validation and cross-sample comparability. (a) Simplified path structure linking treatment, partisan identification, latent misinformation belief, and partisan rating in the experimental sample. (b) Parallel analysis of the four misinformation items, showing clean two-factor retention. (c) Two-factor oblimin loading pattern separating in-party and out-party misinformation items. (d) Kernel density comparison of affective polarization, political sophistication, and ideological extremity across the Social Media and Time Series/ANES samples, with Kolmogorov–Smirnov D and Cohen’s d statistics.

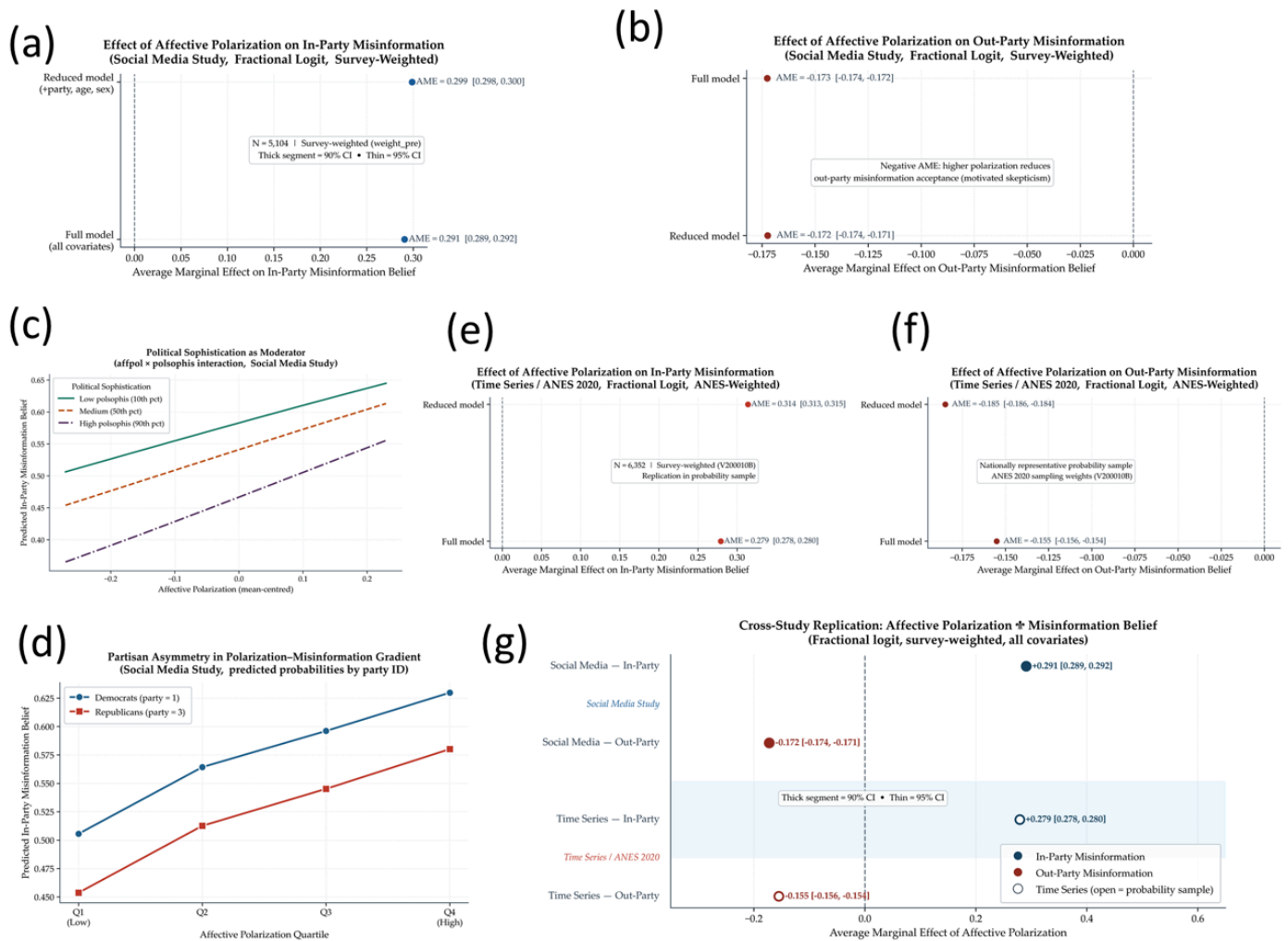


**Figure 3**

Robustness of observational effects across model specifications. (a–b) Social Media in-party and out-party AMEs under reduced versus fully adjusted covariate sets. (c) Predicted in-party misinformation belief as a function of affective polarization at low, medium, and high political sophistication. (d) Predicted in-party misinformation belief by affective-polarization quartile, separately for Democrats and Republicans. (e–f) Time Series in-party and out-party AMEs under reduced versus fully adjusted covariate sets. (g) Cross-study replication summary of all four AMEs.

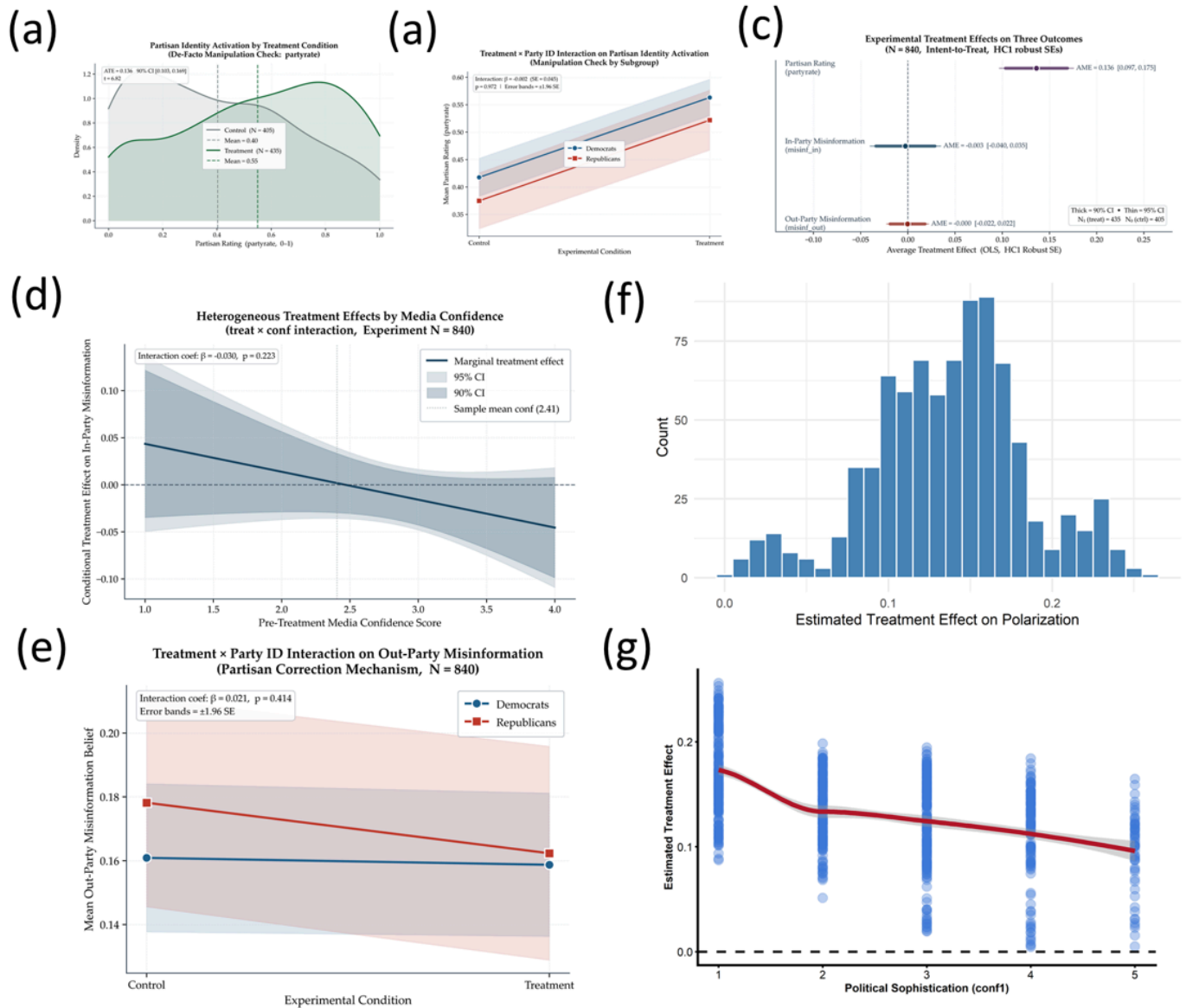


**Figure 4**

Experimental manipulation check and treatment effects. (a) Partisan-identity activation (partyrate) by treatment condition, overall and by party identification. (c) Average treatment effects on partyrate, in-party misinformation, and out-party misinformation. (d) Conditional treatment effect on in-party misinformation as a function of pre-treatment media confidence. (e) Treatment  $\times$  party-identification interaction on out-party misinformation. (f) Distribution of causal-forest individual-level treatment-effect estimates on partisan polarization. (g) Causal-forest estimated treatment effect by political sophistication.

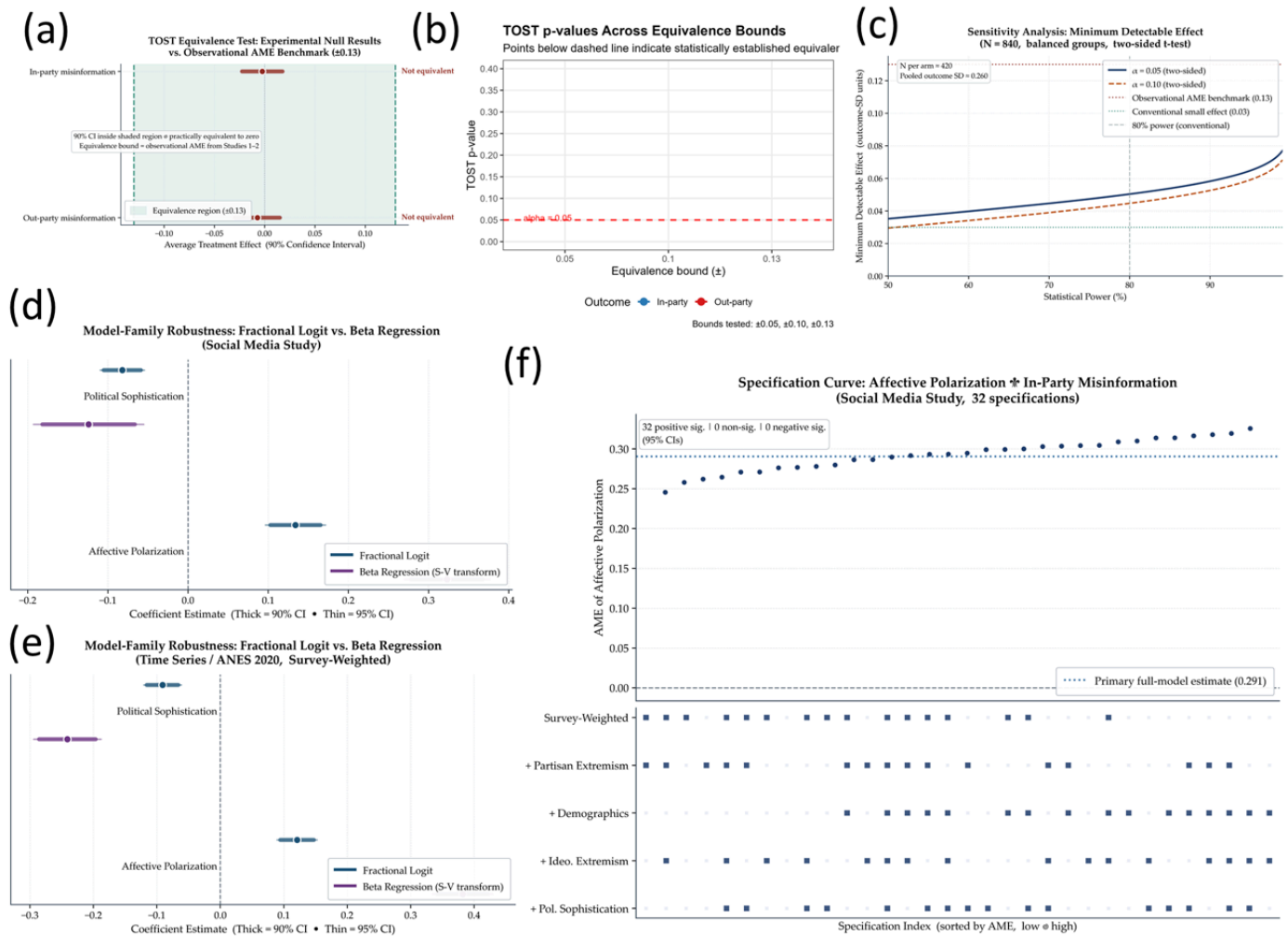


**Figure 5**

Equivalence testing and multiverse robustness analysis. (a) TOST equivalence test of experimental null results against the  $\pm 0.13$  observational benchmark. (b) TOST p-values across a range of equivalence bounds. (c) Minimum detectable effect versus statistical power at  $N = 840$ . (d–e) Fractional logit versus beta-regression coefficient estimates, Social Media and Time Series samples. (f) Specification curve across 32 analytic variants of the Social Media in-party model.